

COVARIATE BALANCING PROPENSITY SCORE BY TAILORED LOSS FUNCTIONS

QINGYUAN ZHAO

Department of Statistics, Stanford University

ABSTRACT. In an observational study, propensity scores are commonly modeled by a generalized linear model (GLM), but the standard maximum likelihood solution may suffer from unsatisfactory covariate balance. This paper proposes to use tailored loss functions—covariate balancing scoring rules (CBSR)—to estimate the propensity score. A CBSR is determined by the link function in the GLM and the estimand (some weighted average treatment effect). If the Bernoulli likelihood criterion is replaced by CBSR to fit the GLM, the resulting inverse probability weights automatically balance all predictors in the GLM. More practical and adaptive strategies are then proposed, including forward stepwise, regularized and reproducing kernel Hilbert space regressions, connecting causal inference with novel machine learning methods. This paper studies the asymptotic efficiency and bias reduction of GLMs fitted by maximizing CBPS, and uses two examples (one simulation, one real data) to study the performance of the proposed methods. Both theoretical and empirical analysis show that CBPS is a superior alternative to Bernoulli likelihood.

1. INTRODUCTION

One major obstacle of obtaining causal relations from observational data is that some pre-treatment covariates are not balanced between the treatment groups. The propensity score, defined as the conditional probability of receiving treatment given the covariates, plays a fundamental role in adjusting for such imbalance. Rosenbaum and Rubin (1983) showed that adjustment by the (scalar) propensity score is sufficient to remove bias due to the potentially high-dimensional covariates. Their seminal work inspired numerous methodological and empirical investigations over the next decades. Today, the propensity score has become an essential tool for applied researchers hunting for causality in observational data.

Most of the propensity-score based methods share three main steps:

Step 1: Estimate a propensity score model, most commonly by maximizing some scoring rule such as the Bernoulli likelihood. A scoring rule is a negative loss function and the two terms will be used interchangeably in this paper. The generalized linear model (McCullagh and Nelder, 1989) has been a workhorse in practice, but more sophisticated alternatives such as nonparametric regression (e.g. Hirano et al., 2003)

E-mail address: qyzhao@stanford.edu.

Date: April 19, 2016.

Key words and phrases. convex optimization, kernel method, inverse probability weighting, proper scoring rule, regularized regression, statistical decision theory, survey sampling.

The author thanks Trevor Hastie, Hera Y. He and Dylan Small for valuable comments.

and machine learning methods (e.g. McCaffrey et al., 2004, Lee et al., 2010, Wager and Athey, 2015) have also been suggested in the literature.

- Step 2:** Adjust for covariate imbalance by using the estimated propensity scores from Step 1. Numerous methods have been proposed, including: matching (e.g. Rosenbaum and Rubin, 1985, Abadie and Imbens, 2006), subclassification (e.g. Rosenbaum and Rubin, 1984), and inverse probability weighting (e.g. Robins et al., 1994, Hirano and Imbens, 2001). The reader is referred to Lunceford and Davidian (2004), Imbens (2004), Caliendo and Kopeinig (2008), Stuart (2010) for some comprehensive reviews.
- Step 3:** Choose a weighted average treatment effect as the estimand and estimate it by using the matches/strata/weights generated in Step 2. Report the point estimate, a confidence interval, evidence of sufficient covariate balance in Step 2 and sensitivity results if necessary.

A leading concern of the propensity-score based methods is that the eventual estimator in Step 3 can be highly sensitive to the outcome of Step 1—the estimated propensity score model (see e.g. Smith and Todd, 2005, Kang and Schafer, 2007). In fact, all the adjustment methods in Step 2 assume that the estimated propensity scores are very close to the truth. This generally requires a correctly specified model or an effective nonparametric regression. In practice, correct model specification is often unrealistic, and nonparametric regression, due to the curse of dimensionality, is a sensible choice only if the sample size is large and the covariates are few. To alleviate the concern of model misspecification, a commonly adopted strategy is to gradually increase the model complexity by forward stepwise regression (Imbens and Rubin, 2015, Section 13.3–13.4). The first two steps described above are usually repeated for several times until satisfactory covariate balance is achieved.

In the standard practice, maximum likelihood is used to fit the propensity score model in Step 1. However, maximum likelihood is suboptimal at balancing covariates. Figure 1 implements the aforementioned forward stepwise strategy with logistic regression and inverse probability weighting (IPW). More detail about this simulation example due to Kang and Schafer (2007) can be found in see Section 6.1. In this Figure, covariate imbalance is measured by the standardized difference of each predictor between the two treatment groups (precise definition in Section 6.2). A widely used criterion is that a standardized difference above 10% is unacceptable (Normand et al., 2001, Austin and Stuart, 2015), which is the dashed line in Figure 1. The left panel of Figure 1 uses the Bernoulli likelihood to fit and select logistic regression models. The standardized difference paths are not monotonically decreasing and never achieve satisfactory level (10%) for more than half of the predictors. This certainly creates inconvenience for applied researchers, and more importantly, limits our understanding of the fundamental bias-variance trade-off in selecting a propensity score model. In contrast, the right panel of Figure 1 uses the covariate balancing scoring rule (CBSR) proposed in this paper and all 8 predictors are well balanced after 4 steps. As another highlighting feature, all active predictors (i.e. variables in the selected model) are exactly balanced using inverse probability weights derived by CBSR.

Why doesn't maximum likelihood always generate covariate balancing weights? Let's review the three-step procedure above, in which the user has the freedom to choose: in Step 1, a form of propensity score model (e.g. certain link function in the GLM) and a scoring rule to fit the model; in Step 2, a propensity-score based adjustment method; in Step 3, a weighted average treatment effect as the estimand. It is understandably tempting to fit a single propensity score model and use it to infer multiple estimands. Along this road, maximum likelihood most efficiently estimates the propensity scores. However, the propensity

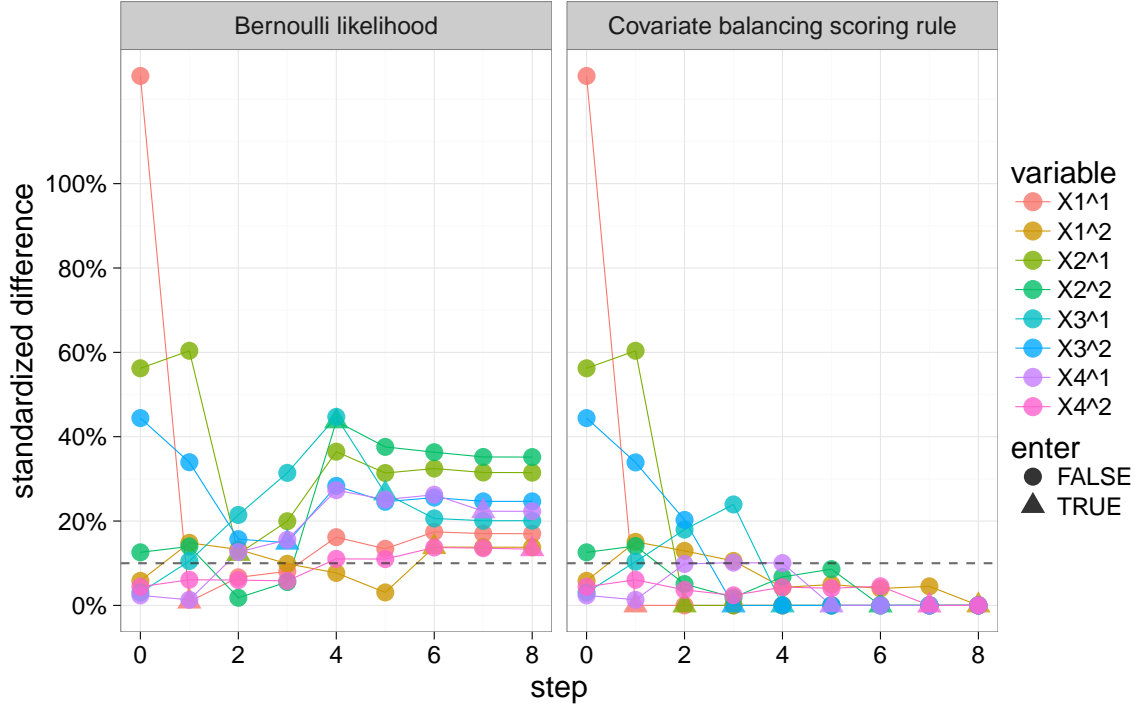


FIGURE 1. The covariate balancing scoring rule (CBSR) proposed in this paper is much better than Bernoulli likelihood at reducing covariate imbalance. Propensity score is modeled by logistic regression and fitted by CBSR or Bernoulli likelihood. Standardized difference is computed using inverse probability weighting (IPW) and pooled variance for the two treatment groups as in Rosenbaum and Rubin (1985), see equation (20) in Section 4.4. A standardized difference above 10% is viewed unacceptable by many practitioners. More detail of the forward stepwise regression and this simulation example can be found in Sections 4.1 and 6.1.

score model is a means to an end, not an end in itself. The most accurate (or even the true) propensity scores do not necessarily produce the best sample balance.

The central message of this paper is that we should tailor the scoring rule according to the estimand, since ultimately we are interested in the estimate in Step 3. CBSR views the three-step procedure as a whole and provides a systematic approach to obtain balancing weights. The CBSR-maximizing propensity scores in Step 1 are best paired with inverse probability weighting (IPW) in Step 2 for its algebraically tractability. As a side note, IPW is also quickly gaining popularity in the literature (Austin and Stuart, 2015) and is more efficient than matching and subclassification. After obtaining the specific form of IPW from the estimand, the covariate balancing score rule can be uniquely determined from the link function of the GLM in Step 1.

1.1. Related Work and Our Contribution. The covariate balancing scoring rule (CBSR) is largely inspired by some recent approaches that directly incorporate covariate balance in propensity score estimation. Imai and Ratkovic (2014) proposed to augment the Bernoulli

likelihood with the covariate balancing estimating equations, hoping they can robustify the propensity score model. These estimating equations are exactly the first-order conditions of maximizing CBSR. However, our derivation of CBSR shows that different estimating equations correspond to different estimands and there is little reason to combine them. In fact, Imai and Ratkovic (2014) also found in their simulation study that just using the covariate balancing estimating equations usually performs better. Another distinction is that Imai and Ratkovic (2014) solved the estimating equations by generalized method of moments or empirical likelihood. Those generic methods are generally not convex. This paper identifies the scoring rules corresponding to the estimating equation (10) and Proposition 2 shows that they are concave for estimating ATE and ATT with the logistic link function. Convex optimization methods can be used to solve the score maximization problem very efficiently.

Another related approach is Hainmueller (2011)’s Entropy Balancing which specializes in estimating ATT. It operates by maximizing the Shannon entropy of sample weights subject to exact covariate balance. Zhao and Percival (2015) found that the Lagrangian dual of Entropy Balancing fits a logistic propensity score model with a loss function different from the Bernoulli likelihood. The present paper generalizes this approach to general estimands.

To summarize, the decision theoretical approach we take has a number of advantages:

1. The scoring rules (loss functions) can be plotted and interpreted visually, showing how propensity score estimation should be treated differently than a standard classification problem. CBSR penalizes more heavily on larger inverse probability weights hence generates a more stable estimator.
2. A proper scoring rule generates Fisher consistent estimates of the propensity score, allowing us to study the asymptotic properties of the IPW estimators.
3. The Lagrangian duality connects IPW estimators with the calibration estimators in survey sampling (Deville and Särndal, 1992). It also demonstrates an explicit bias-variance trade-off with regularized propensity score models.
4. The convex loss function opens up numerous opportunities to use machine learning algorithms to estimate the propensity score. These algorithms are usually designed to optimize predictive performance. With the covariate balancing scoring rules as the objective, the machine learning algorithms now try to optimize covariate balance between the treatment groups. Several extensions are discussed in Section 4 and other potential extensions such as boosting, decision trees and Bayesian methods will be pursued in future work.

The next three Sections are devoted to introducing the covariate balancing scoring rules and some practical strategies motivated by machine learning. Section 5 then considers some theoretical aspects about CBSR. Section 6 uses two numerical examples to show CBSR has superior empirical performance than the Bernoulli likelihood.

2. BACKGROUND ON STATISTICAL DECISION THEORY

Propensity score estimation is a decision problem, though an unusual one. In a typical decision problem of making probabilistic forecast, the decision maker needs to pick an element as the prediction from \mathcal{P} , a convex class of probability measures on some general sample space Ω . For example, a weather forecaster needs to report the chance of rain tomorrow, so the sample space is $\Omega = \{\text{rain, no rain}\}$ and the prediction is a Bernoulli distribution. Propensity score is also a (conditional) probability measure, but the goal is to achieve satisfactory covariate balance rather than best predictive power. This marks a clear difference to the

prediction problem. Nevertheless, statistical decision theory provides a general framework and effective tools to fit a covariate balancing propensity score model.

2.1. Proper scoring rules. Let's first review some useful concepts. At the core of statistical decision theory is the *scoring rule*, which can be any extended real-valued function $S : \mathcal{P} \times \Omega \rightarrow [-\infty, \infty]$ such that $S(P, \cdot)$ is \mathcal{P} -integrable for all $P \in \mathcal{P}$ (Gneiting and Raftery, 2007). If the decision is P and ω materializes, the decision maker's reward or utility is $S(P, \omega)$. An equivalent but more pessimistic terminology is *loss function*, which is just the negative scoring rule. These two terms will be used interchangeably in this paper.

If the outcome is probabilistic in nature and the actual probability distribution is Q , the expected score of forecasting P is

$$S(P, Q) = \int S(P, \omega) Q(d\omega).$$

To encourage honest decisions, we generally require the scoring rule S to be *proper* with respect to \mathcal{P} that is defined by

$$S(Q, Q) \geq S(P, Q), \quad \forall P, Q \in \mathcal{P}. \quad (1)$$

The rule is called *strictly proper* with respect to \mathcal{P} if (1) holds with equality if and only if $P = Q$. In estimation problems, strictly proper scoring rules provide appealing loss functions that can be tailored according to the scientific problem.

In observational studies, the sample space is commonly dichotomous $\Omega = \{0, 1\}$ (two treatment groups: 0 for control and 1 for treated), though there is no essential difficulty to extend the approach in this paper to $|\Omega| > 2$ (multiple treatments) or $\Omega \subset \mathbb{R}$ (continuous treatment). In the binary case, Savage (1971) showed that if $S(\cdot, 0)$ and $S(\cdot, 1)$ are real-valued except for possibly $S(0, 1) = \infty$ or $S(1, 0) = -\infty$, every proper scoring rule S can be characterized by

$$\begin{aligned} S(p, 1) &= G(p) + (1 - p)G'(p) = \int (1 - p)G''(p)dp, \\ S(p, 0) &= G(p) - pG'(p) = - \int pG''(p)dp, \end{aligned}$$

where $G : [0, 1] \rightarrow \mathbb{R}$ is a convex function and $G'(p)$ is a subgradient of G at the point $p \in [0, 1]$. When G is second-order differentiable, an equivalent but useful representation is

$$\frac{\partial}{\partial p} S(p, t) = (t - p)G''(p), \quad t = 0, 1. \quad (2)$$

Since the function G defines an equivalent class of scoring rule, we shall also call G a scoring rule.

A useful class of proper scoring rules is the following Beta family

$$G''_{\alpha, \beta}(p) = p^{\alpha-1}(1-p)^{\beta-1}, \quad -\infty < \alpha, \beta < \infty. \quad (3)$$

These scoring rules were first introduced by Buja et al. (2005) to approximate the weighted misclassification loss by taking the limit $\alpha, \beta \rightarrow \infty$ and $\alpha/\beta \rightarrow c$. For example, if $c = 1$, the score $G_{\alpha, \beta}$ converges to the zero-one misclassification loss. Many important scoring rules belong to this family. For example, the Bernoulli log-likelihood function or the logarithmic score $S(p, t) = t \log p + (1 - t) \log(1 - p)$ corresponds to $\alpha = \beta = 0$, and the Brier score (or equivalently the squared error loss when flipping the sign) $S(p, t) = -(t - p)^2$ corresponds to

$\alpha = \beta = 1$. For our purpose of estimating propensity score, it will be shown later that the subfamily $-1 \leq \alpha, \beta \leq 0$ is especially useful.

2.2. Propensity score estimation by maximizing score. Given i.i.d. observations $(X_i, T_i) \in \mathbb{R}^d \times \{0, 1\}$, $i = 1, 2, \dots, n$ where T_i is the binary treatment assignment and X_i is a vector of d pre-treatment covariates, the goal is to fit a model for the propensity score $p(X) = P(T|X)$. Suppose we are willing to use a parametric model that belongs to the family $\mathcal{P} = \{p_\theta(X) : \theta \in \Theta\}$. Given a strictly proper scoring rule S , the goodness-of-fit of a given θ can be measured by the average score

$$\mathcal{S}_n(\theta) = \frac{1}{n} \sum_{i=1}^n S(p_\theta(X_i), T_i),$$

The *optimum score estimator* is obtained by the unique maximizer of the average score:

$$\hat{\theta}_n = \arg \max_{\theta} \mathcal{S}_n(\theta) \quad (4)$$

Notice that the affine transformation $S(p, t) \mapsto aS(p, t) + b(t)$ for any $a > 0$ and $-\infty < b(t) < \infty$ results in the same estimator $\hat{\theta}_n$, so we shall not differentiate between these equivalent scoring rules and use a single function $S(p, t)$ to represent all equivalent ones.

In view of the population identity

$$E[\mathcal{S}_n(\theta)] = E_{X,T}[S(p_\theta(X), T)] = E_X[E_{T|X}[S(p_\theta(X), T)]],$$

the optimum score estimator is Fisher-consistent. Fisher-consistency means that the true value of the parameter θ would be obtained if the true propensity score is $p(x) = p_\theta(x)$ and the estimator were calculated using the entire population rather than a sample. In many cases, including the problem considered in this paper, this property also leads to asymptotic consistency: $\hat{\theta}_n \xrightarrow{P} \theta$ as $n \rightarrow \infty$.

This paper focuses on using the generalized linear models (McCullagh and Nelder, 1989)

$$p_\theta(X) = l^{-1}(f_\theta(X)) = l(\theta^T \phi(X)) \quad (5)$$

to model the propensity score. Here l is the *link function*, $f_\theta(X)$ is the *canonical parameter* which is modeled by a linear combination of the m -dimensional *predictors* $\phi(X) = (\varphi_1(X), \dots, \varphi_m(X))^T$. The covariate balancing scoring rule derived in this paper depends on the link function l . The most common choice is the logistic link:

$$l(p) = \log \frac{p}{1-p}, \quad l^{-1}(f) = \frac{e^f}{1+e^f}. \quad (6)$$

This will be the choice for all the numerical examples in this paper.

When S is differentiable and assuming exchangeability of taking expectation and derivative, the maximizer of $E[\mathcal{S}_n(\theta)]$, which is indeed θ if $p(x) = p_\theta(x)$ by Fisher-consistency, is characterized by the following estimating equations

$$\nabla_\theta E[\mathcal{S}_n(\theta)] = E[\nabla_\theta \mathcal{S}_n(\theta)] = E_{X,T}[\nabla_\theta S(l^{-1}(\theta^T \phi(X)), T)] = 0. \quad (7)$$

Using the representation (2) and the inverse function theorem, we have

$$\nabla_\theta S(l^{-1}(\theta^T \phi(X)), T) = (T - p_\theta(X))G''(p_\theta(X)) \frac{1}{l'(p_\theta(X))} \cdot \phi(X).$$

Therefore the condition (7) can be written as

$$E_{X,T} \left\{ \frac{G''(p_\theta(X))}{l'(p_\theta(X))} [T(1 - p_\theta(X)) - (1 - T)p_\theta(X)] \cdot \phi(X) \right\} = 0. \quad (8)$$

The optimum score estimator, $\hat{\theta}_n$, can be determined from (8) by taking the expectation over the empirical distribution of (X, T) , provided that S is strictly proper so the solution to (8) is unique.

3. COVARIATE BALANCING SCORING RULES

The covariate balancing scoring rules (CBSR) are motivated by the estimating equations (8), which can be interpreted as weighted differences of $\phi(X)$ between the treatment ($T = 1$) and the control ($T = 0$). The weights are given by, for $t = 0, 1$,

$$w(x, t) = \frac{G''(p(x))}{l'(p(x))} [t(1 - p(x)) + (1 - t)p(x)]. \quad (9)$$

Equation (8) can now be rewritten as stochastic balance of the predictors

$$E[(T - (1 - T))w(X, T) \cdot \phi(X)] = 0, \quad (10)$$

The question is: Are these weights meaningful? In other words, do they correspond to some form of inverse probability weighting (IPW)?

3.1. Covariate balancing scoring rules. The answer to this question is, of course, positive. In short, every convex function G defines a weighted average treatment effect via (9). To see this we need to introduce some notation. Following the Neyman-Rubin causal model, let $Y(t)$, $t = 0, 1$ be the potential outcomes and $Y = TY(1) + (1 - T)Y(0)$ be the observed outcome. This paper assumes the following strong ignorability of treatment assignment (Rosenbaum and Rubin, 1983), so there is no hidden bias:

Assumption 1. $T \perp (Y(0), Y(1)) | X$.

First, we define a population parameter by replacing $\phi(X)$ in (10) with the outcome Y

$$\tau_w = E_{X,T,Y} \{(T - (1 - T))w(X, T)Y\},$$

Under Assumption 1, τ_w is indeed an (unnormalized) weighted average treatment effect

$$\tau_w = E_{X,Y} [w(X)(Y(1) - Y(0))], \quad (11)$$

where

$$w(X) = p(X)w(X, 1) = (1 - p(X))w(X, 0) = \frac{G''(p(X))p(X)(1 - p(X))}{l'(p(X))}.$$

In practice, it is usually more meaningful to consider the normalized version of τ_w :

$$\tau_w^* = \tau_w \bigg/ E_{X,T,Y} \left[\frac{G''(p(X))p(X)(1 - p(X))}{l'(p(X))} \right]. \quad (12)$$

The question now becomes: is τ_w^* an interesting estimand in observational studies? The answer to this question is, again, positive. Consider the following Beta family of weighted average treatment effects

$$\tau_{\alpha,\beta} = E[p(X)^{\alpha+1}(1 - p(X))^{\beta+1}(Y(1) - Y(0))], \quad -1 \leq \alpha, \beta \leq 0. \quad (13)$$

Several important estimands belong to this family, including the average treatment effect (ATE), the average treatment effect on the untreated (ATUT), the average treatment effect on the treated (ATT), and the optimally weighted average treatment effect under homoscedasticity (Crump et al., 2006). See the third column of Table 1 for the definitions of these estimands.

α	β	estimand	$S(p, 1)$	$S(p, 0)$
-1	-1	$\tau = \tau^* = \mathbb{E}[Y(1) - Y(0)]$	$\log \frac{p}{1-p} - \frac{1}{p}$	$\log \frac{1-p}{p} - \frac{1}{1-p}$
-1	0	$\tau^* = \mathbb{E}[Y(1) - Y(0) T = 0]$	$-\frac{1}{p}$	$\log \frac{1-p}{p}$
0	-1	$\tau^* = \mathbb{E}[Y(1) - Y(0) T = 1]$	$\log \frac{p}{1-p}$	$-\frac{1}{1-p}$
0	0	$\tau = \mathbb{E}[p(X)(1 - p(X)) \cdot (Y(1) - Y(0))]$	$\log p$	$\log(1 - p)$

TABLE 1. Estimands and scoring rules in the Beta family.

The next Proposition shows an exact correspondence between the Beta family of estimands (13) and the Beta family of scoring rules (3).

Proposition 1. *Under Assumption 1, if $G = G_{\alpha, \beta}$ and l is the logistic link function, then $\tau_w = \tau_{\alpha, \beta}$.*

Proof. Use equations (3), (6), (9), (11) and (13). \square

Therefore, some of the most important estimands in observational studies can be defined by (12). Proposition 1 also suggests a general strategy to estimate average causal effect:

1. Pick a weighted average treatment effect $\tau = \tau_{\alpha, \beta}$ from the Beta family (13) as the estimand.
2. Compute its corresponding scoring rule using (9) or find it from Table 1 below.
3. Using the scoring rule, fit a logistic regression $\hat{p}(X) = l^{-1}(\hat{\theta}^T \phi(X))$ for the propensity score.
4. Estimate τ and its normalized version τ^* defined in (12) by

$$\hat{\tau} = \sum_{i: T_i=1} \hat{w}_i Y_i - \sum_{i: T_i=0} \hat{w}_i Y_i \text{ and } \hat{\tau}^* = \sum_{i: T_i=1} \hat{w}_i^* Y_i - \sum_{i: T_i=0} \hat{w}_i^* Y_i, \quad (14)$$

where

$$\hat{w}_i = p_{\hat{\theta}}(X_i)^\alpha (1 - p_{\hat{\theta}}(X_i))^\beta [T_i(1 - p_{\hat{\theta}}(X_i)) + (1 - T_i)p_{\hat{\theta}}(X_i)] \quad (15)$$

and the normalized weights are $\hat{w}_i^* = \hat{w}_i / \sum_{j: T_j=T_i} \hat{w}_j$, $i = 1, \dots, n$.

A main advantage of this approach is that the weights automatically balance the predictors $\phi(X)$ in the logistic regression, as indicated by the next theorem.

Theorem 1. *Given a scoring rule $S_{\alpha, \beta}$ in the Beta family and a logistic regression model $p_{\theta}(X) = l^{-1}(\theta^T \phi(X))$, suppose $\hat{\theta}$ is obtained by maximizing the average score as in (4). Then the weights \hat{w}_i , $i = 1, \dots, n$, exactly balance the sample predictors*

$$\sum_{i: T_i=1} \hat{w}_i \phi(X_i) = \sum_{i: T_i=0} \hat{w}_i \phi(X_i). \quad (16)$$

Furthermore, if the predictors include an intercept term (i.e. 1 is in the linear span of $\phi(X)$), then \hat{w}^* also satisfies (16).

Proof. This theorem is a simple corollary of the estimating equations (10). \square

Because of Theorem 1, $G_{\alpha,\beta}$ or the resulting $S_{\alpha,\beta}$ will be called the *covariate balancing scoring rule* (CBSR) with respect to the estimand $\tau_{\alpha,\beta}$ and the logistic link function.

3.2. A closer look at the Beta family. One may wonder why the estimands in (13) are restricted to the subfamily $-1 \leq \alpha, \beta \leq 1$. There are at least two reasons. First, as mentioned earlier, this sub-family already contains most of the important estimands that are meaningful to observational studies (see Table 1). Second, as shown in Proposition 2 below, this is the only region such that the maximum score problem (4) is convex when $p_\theta(X)$ is modeled by logistic regression. Therefore the optimization problem (4) has no local maximum and can be solved efficiently (e.g. by Newton’s method).

Proposition 2. *For the Beta family of scoring rules defined in equations (2) and (3) and the logistic link function $l^{-1}(f) = e^f / (1 + e^f)$, the score functions $S(l^{-1}(f), 0)$ and $S(l^{-1}(f), 1)$ are both concave functions of $f \in \mathbb{R}$ if and only if $-1 \leq \alpha, \beta \leq 1$. Moreover, if $(\alpha, \beta) \neq (-1, 0)$, $S(l^{-1}(f), 0)$ is strongly concave; if $(\alpha, \beta) \neq (0, -1)$, $S(l^{-1}(f), 1)$ is strongly concave.*

Proof. See Appendix A.1. □

Figure 2 plots the scoring rules $S_{\alpha,\beta}$ for some combinations of α and β . The top panels show the score function $S(p, 0)$ and $S(p, 1)$ for $0 < p < 1$, which are normalized so that $S(1/4, 1) = S(3/4, 0) = -1$ and $S(1/4, 0) = S(3/4, 1) = 1$. By a change of variable, one can show $S_{\alpha,\beta}(p, 1) = S_{\beta,\alpha}(1 - p, 0)$. This is the reason that the two subplots in Figure 2a are essentially reflections of each other. The bottom panels show the induced scoring rule $S(p, q)$ defined by section 2.1 or more specifically $S(p, q) = qS(p, 1) + (1 - q)S(p, 0)$ at two different values of $q = 0.05, 0.15$. For aesthetic purposes, the scoring rules in Figure 2b are normalized such that $-S(p, q) = 1$ and $-S(p, 1 - q) = 2$.

Figure 2 shows that the scoring rules $S_{\alpha,\beta}$, when $-1 \leq \alpha, \beta \leq 0$, are highly sensitive to small differences of small probabilities. For example, in Figure 2a the loss function $-S_{\alpha,\beta}(p, 1)$ is unbounded above when $\alpha, \beta \in \{-1, 0\}$, hence a small change of p near 0 may have a big impact on the score. In Figure 2b, the averaged scoring rules $S_{\alpha,\beta}(p, q)$, when $(\alpha, \beta) = (-1, -1)$ or $(-1, 0)$, are also unbounded near $p = 0$. Due to this reason, Selten (1998, Section 2.6) argued that these scoring rules are inappropriate for probability forecast problems.

On the contrary, the unboundedness is actually a desirable feature for propensity score estimation, as the goal is to avoid extreme probabilities. Consider the standard inverse probability weights (IPW)

$$\hat{w}_i = \begin{cases} \hat{p}_i^{-1} & \text{if } T_i = 1, \\ (1 - \hat{p}_i)^{-1} & \text{if } T_i = 0, \end{cases} \quad (17)$$

where $\hat{p}_i = p_{\hat{\theta}}(X_i)$ is the estimated propensity score for the i -th data point. This corresponds to $\alpha = \beta = -1$ in the Beta family and estimates ATE. Several previous articles (e.g. Robins and Wang, 2000, Kang and Schafer, 2007, Robins et al., 2007) have pointed out the hazards of using large inverse probability weights. For example, if the true propensity score is $p(X_i) = q = 0.05$ and it happens that $T_i = 1$, we would want \hat{p}_i not too close to 0 so \hat{w}_i is not too large. Conversely, we also want \hat{p}_i not too close to 1, so in the more likely event that $T_i = 0$ the weight \hat{w}_i is not too large either. In an *ad hoc* attempt to mitigate this issue, Lee et al. (2011) studied weight truncation (e.g. truncate the largest 10% weights). They found that the truncation can reduce the standard error of the estimator $\hat{\tau}$ but also increases the bias.

The covariate balancing scoring rules provide a more systematic approach to avoid large weights. For example, the scoring rule $S_{-1,-1}$ precisely penalizes large inverse probability

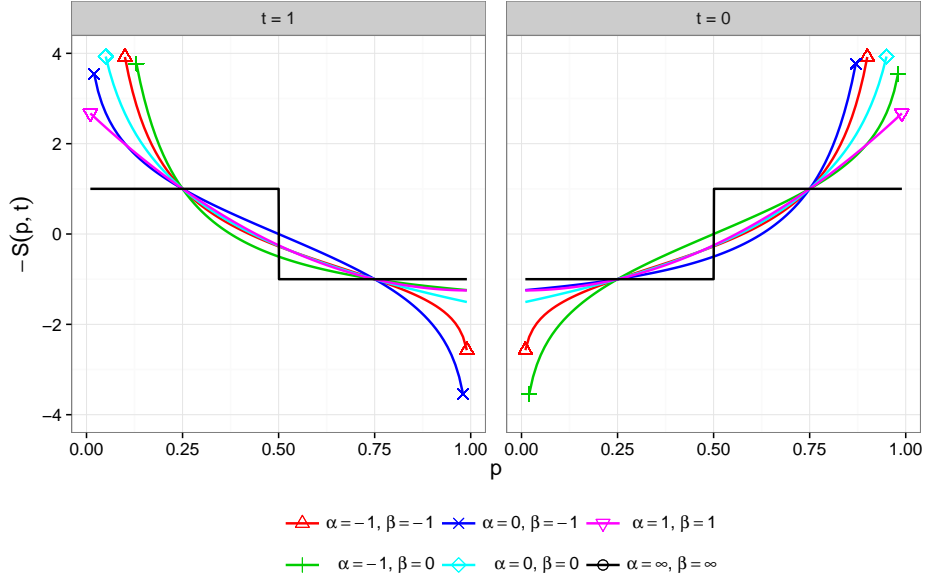
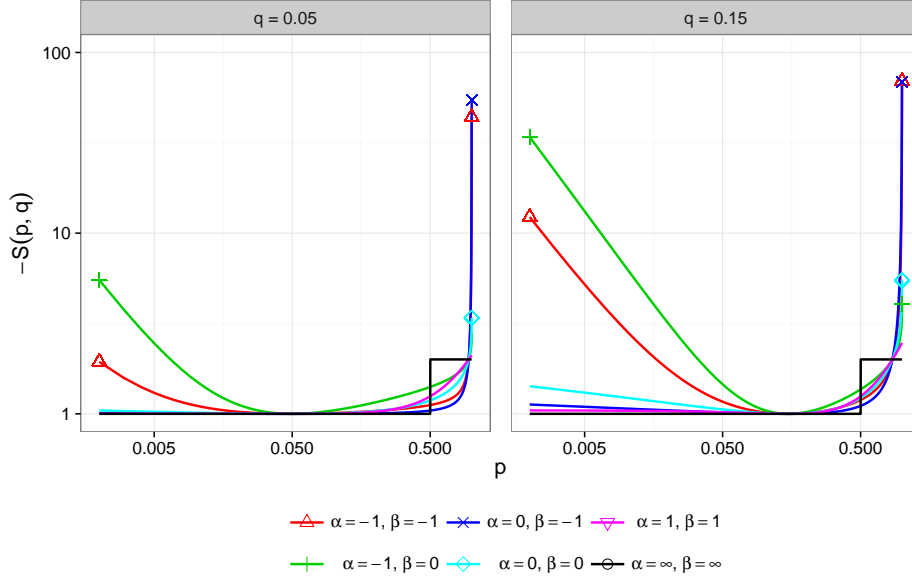
(A) Loss functions $-S_{\alpha,\beta}(p, t)$ for $t = 0, 1$.(B) Loss functions $-S_{\alpha,\beta}(p, q)$ for $q = 0.05$ and 0.15 .

FIGURE 2. Graphical illustration of the Beta-family of scoring rules defined in (3).

weights as $-S_{-1,-1}(p, q)$ is unbounded above when p is near 0 or 1 (see the left plot in Figure 2b). Similarly, when estimating the ATUT $\tau_{-1,0}$, the weighting scheme would put $\hat{w}_i \propto (1 - \hat{p}_i)/\hat{p}_i$ if $T_i = 1$ and $\hat{w}_i \propto 1$ if $T_i = 0$. Therefore we would like \hat{p}_i to be not close to 0, but it is acceptable if \hat{p}_i is close to 1. As shown in in Figure 2b, the curve

$-S_{-1,0}(p, q) = q/p + (1-q) \log(p/(1-p))$ precisely encourages this behavior, as it is unbounded above when p is near 0 and grows slowly when p is near 1.

4. ADAPTIVE STRATEGIES

So far we have only considered a fixed GLM to model the propensity score. This Section discusses some adaptive extensions motivated by popular machine learning algorithms. In order to achieve the best predictive performance, most machine learning methods prespecify a loss function to train the model. For the purpose of obtaining covariate balancing weights, we only need to replace the loss function by the covariate balancing scoring rule (CBSR) introduced in this paper. This is indeed a major advantage of using scoring rules instead of estimating equations.

4.1. Forward Stepwise. Let's start with the forward stepwise regression which is already widely used in observational studies (Imbens and Rubin, 2015). The notation $\phi(x) = (\phi_1(x), \dots, \phi_m(x))$ is used to indicate all the potential linear predictors. This entire Section allows $m > n$, but it is not necessary to include all the predictors in the model.

Algorithm 1 Forward stepwise regression for propensity score

Input data: (T_i, X_i) , $i = 1, \dots, n$.

Input arguments: predictors $\{\phi_1(x), \dots, \phi_m(x)\}$, link function $l(\cdot)$, proper scoring rule $S(p, t)$.

Notation: $\mathcal{F}_{\mathcal{A}} = \text{span}(\{\phi_k(x) | k \in \mathcal{A}\})$.

Algorithm:

Initialize active set $\mathcal{A} = \emptyset$.

for $j = 1, \dots, m$ **do**

Compute $S_{jk} = \max_{f \in \mathcal{F}_{\mathcal{A} \cup \{k\}}} \sum_{i=1}^n S(l^{-1}(f(X_i)), T_i)$ for $k \in \mathcal{A}^c$.

Update $\mathcal{A}_k = \mathcal{A} \cup \{\arg \max_k S_{jk}\}$.

end for

Output:

\mathcal{A}^* from \mathcal{A}_k , $k = 1, \dots, m$ that optimizes some criterion (e.g. AIC, BIC, least covariate imbalance).

$f^* = \arg \max_{f \in \mathcal{F}_{\mathcal{A}^*}} \sum_{i=1}^n S(l^{-1}(f(X_i)), T_i)$.

In Algorithm 1, the predictors are added one by one in a forward stepwise regression. After choosing a scoring rule, the algorithm in each step fits a GLM using all the selected predictors and each unselected predictor. The unselected predictor that increases the score \mathcal{S}_n the most is added to the active set. This procedure is repeated until no new predictor can be added or the current score \mathcal{S}_n is already ∞ . Figure 1 demonstrates this adaptive algorithm with a simulation example described in Section 6.1. There is no need to reiterate that CBSR is much better at reducing covariate imbalance than Bernoulli likelihood.

4.2. Regularized Regression. Another widely-used adaptive method is the following regularized solution of the GLM (5):

$$\hat{\theta}_\lambda = \arg \max_{\theta} \frac{1}{n} \sum_{i=1}^n S(p_\theta(X_i), T_i) - \lambda J(\theta), \quad (18)$$

where $J(\cdot)$ is a regularization term that penalizes large θ (complicated model) and λ controls the degree of regularization. This estimator reduces to the optimum score estimator (4) when $\lambda = 0$. For simplicity, this paper only considers penalty of the form

$$J(\theta) = \frac{1}{a} \sum_{k=1}^m |\theta_k|^a \text{ for some } a \geq 1. \quad (19)$$

Some typical choices are the l_1 norm $J(\theta) = \|\theta\|_1$ (lasso) and the squared l_2 norm $J(\theta) = \|\theta\|_2^2$ (ridge regression).

An important advantage of the regularized regression (18) is that it allows high dimensional predictors $\phi(X)$. This is useful to propensity score estimation for at least three reasons:

1. The pre-treatment covariates X can be high dimensional, especially if we wish to follow Rubin (2009)’s advice that “we should strive to be as conditional as is practically possible”.
2. Even if X is relatively low dimensional, we may still want to use a high dimensional $\phi(X)$ to essentially build a nonparametric propensity score model.
3. The Beta family of scoring rules (3) with $-1 \leq \alpha, \beta \leq 0$ are unbounded above, so $\sup_{\theta} \mathcal{S}_n(\theta)$ can easily be infinity if ϕ is high dimensional, making the optimum score problem (4) infeasible. The Bernoulli likelihood ($\alpha = \beta = 0$) also suffers from this. In this case, it is necessary to add some regularization as in (18) to obtain any propensity score model.

4.3. Kernel method. The predictors $\phi(X)$ can even be infinite dimensional via a popular nonparametric regression method in machine learning (Wahba, 1990, Hofmann et al., 2008, Hastie et al., 2009). This method models the propensity score $p(x)$ by $l^{-1}(f(x))$ with f in a reproducing kernel Hilbert space (RKHS) \mathcal{H}_K , where the kernel function $K : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ describes the similarity between two observations of pre-treatment covariates. Suppose that K has an eigen-expansion

$$K(x, x') = \sum_{k=1}^{\infty} c_k \phi_k(x) \phi_k(x')$$

with $c_k \geq 0$, $\sum_{k=1}^{\infty} c_k^2 < \infty$. Elements of \mathcal{H}_K have an expansion in terms of these eigen-functions,

$$f(x) = \sum_{k=1}^{\infty} \theta_k \phi_k(x).$$

The standard generalized linear model (5) corresponds to a finite-dimensional linear reproducing kernel $K(x, x') = \sum_{k=1}^m \phi_k(x) \phi_k(x')$, but the eigen-functions (i.e. predictors) $\{\phi_k\}_{k=1}^{\infty}$ can easily be infinite-dimensional. Since the RKHS is usually a very rich function space, it is common to regularize the score as in (18) with penalty $J(\theta) = \|f\|_{\mathcal{H}_K}^2 = \sum_{k=1}^{\infty} \theta_k^2 / c_k$.

Although RKHS incorporates potentially infinite-dimensional predictors, the numerical problem (18) is computationally feasible via the so-called “kernel trick”. The representer theorem (c.f. Wahba, 1990) states that the solution to (18) is indeed finite-dimensional and has the form $\hat{f}(x) = \sum_{i=1}^n \hat{\gamma}_i K(x, X_i)$. Consequently, the optimization problem (18) can be solved with the n -dimension parameter vector γ .

Kernels are not newcomers to the toolbox for observational studies. Most of the previous literature (e.g. Heckman et al., 1997, 1998) uses kernel as a smoothing technique for propensity score estimation (i.e. a generalization of the nearest neighbor matching) rather than

generating a RKHS, but the kernel function K is the same. The tuning parameters in RKHS are the kernel bandwidths and the amount of smoothness penalty. Sensitivity analysis may be carried out with little extra effort by varying over different kernel forms and bandwidths.

The RKHS approach has another practical benefit: the modeling process is free from guessing model specifications. The user only needs to choose a kernel that measures the closeness of two units based on pre-treatment covariates X . It is arguably much easier for a field expert to answer questions like “is patient A or patient B more similar to patient C based on their age and education?” than to speculate and make sense of a model like “the logit of the propensity score is linear in age and years of education”.

4.4. Model selection and inference. After a series of propensity score models are fitted by forward stepwise regression or regularized regression as described earlier, the remaining question is to select one model for further statistical inference. This is a standard task in propensity-score based approaches. The selected model should best balance all the pre-treatment covariates. One measurement of covariate imbalance is the absolute standardized difference (Rosenbaum and Rubin, 1985). For the estimated weights \hat{w} and each predictor ϕ_k , $k = 1, \dots, m$, it is defined as

$$d_k = \frac{\left| (1/n_1) \sum_{i:T_i=1} \hat{w}_i \phi_k(X_i) - (1/n_0) \sum_{j:T_j=0} \hat{w}_j \phi_k(X_j) \right|}{s_w}, \quad (20)$$

where s_w^2 is the sample variance of the numerator in (20). We can also use the t -test based on (20) to verify the means are not significantly different. Another widely used criterion is the nonparametric Kolmogorov-Smirnov test. The reader is referred to the review articles by Caliendo and Kopeinig (2008), Austin and Stuart (2015) for more practical guidance. In the simulation example in Section 6.1, we choose the propensity score model that has the smallest number of significant two-sample t -tests. When the covariate balancing scoring rule is used, the selected model is usually close to the end of the path.

Given a propensity score model, the weighted average treatment effect τ or its normalized τ^* can be estimated by inverse probability weighting described in (14) and (15). To obtain a confidence interval for τ or τ^* , we adopt a general method for estimating sampling variances in Imbens and Rubin (2015, Chapter 19). Let $\text{Var}(Y_i) = \sigma_i^2$. Conditioning on the covariates X and the estimated weights \hat{w} , the sampling variance of $\hat{\tau}$ is given by

$$\text{Var}(\hat{\tau}|X, w) = \sum_{i=1}^n \hat{w}_i^2 \sigma_i^2. \quad (21)$$

Imbens and Rubin (2015, Section 19.6) described several ways to estimate σ_i^2 for all units. In the numerical examples in Section 6, we assume additive homoskedastic noise $\sigma_i^2 = \sigma^2$ and use a pilot outcome regression to estimate the noise variance σ^2 .

5. THEORETICAL ASPECTS

This Section discusses the following four theoretical aspects about CBSR. A first-time reader more interested in the empirical performance can skip the next two Sections and go to the numerical examples in Section 6.

1. With increasingly complex propensity score model as the sample size grows, any strongly concave proper scoring rule can provide semiparametrically efficient estimate of the weighted average treatment effect (Section 5.1).

2. Even if the propensity score model is misspecified, CBSR can still reduce the bias and the variance of $\hat{\tau}$ due to the covariate balancing weights (Section 5.2).
3. The Lagrangian dual of the CBSR maximization problem is an entropy maximization problem with covariate balancing constraints. This observation connects IPW estimators with calibration estimators in survey sampling (Section 5.3).
4. The Lagrangian duality also allows us to study the bias-variance tradeoff in selecting propensity score models (Section 5.4).

5.1. Global efficiency by sieve regression. If a statistician is asked about why maximum likelihood is the predominantly used scoring rule, most likely he/she will refer to its attractive limiting properties—consistency, asymptotic normality, and most importantly, efficiency, i.e. maximum likelihood can reach the Cramér-Rao bound. However, as mentioned in Section 1, the ultimate goal in an observational study is to infer some average treatment effect. Propensity score model, no matter fitted by maximizing Bernoulli likelihood or CBSR, is just a means to this end. A natural question is: is it necessary or even beneficial to fit the propensity score model most efficiently by maximum likelihood?

Here we study the efficient estimation of weighted average treatment effects in the setting of nonparametric sieve regression. As the sample size n grows, a *sieve* estimator uses progressively more complex models to estimate the unknown propensity score. For example, we can increase the dimensional of the predictors in $\phi(x)$ in the GLM (5). This approach is used in Hirano, Imbens, and Ridder (2003) to estimate the propensity score by maximum likelihood. Their renowned results claim that the resulting IPW estimator is globally efficient for estimating ATE, ATT and other weighted average treatment effects. It is shown below that the global efficiency still holds if the Bernoulli likelihood is changed to the Beta family of scoring rules $G_{\alpha,\beta}$, $-1 \leq \alpha, \beta \leq 0$ in (3) or essentially any strongly concave scoring rule. Therefore there is no efficiency gain by sticking to the likelihood criterion.

First, let's briefly review the sieve logistic regression in Hirano et al. (2003). For $m = 1, 2, \dots$, let $\phi_m(x) = (\varphi_{1m}(x), \varphi_{2m}(x), \dots, \varphi_{mm}(x))^T$ be a triangular array of orthogonal polynomials, which are obtained by orthogonalizing the power series: $\psi_{km}(x) = \prod_{j=1}^d x_j^{\gamma_{kj}}$, where $\gamma_k = (\gamma_{k1}, \dots, \gamma_{kd})^T$ is an d -dimensional multi-index of nonnegative integers and satisfies $\sum_{j=1}^d \gamma_{kj} \leq \sum_{j=1}^d \gamma_{k+1,j}$. Let l be the logistic link function (6). Hirano et al. (2003) estimated the propensity score by the following maximum likelihood rule

$$\hat{\theta}^{\text{MLE}} = \arg \max_{\theta} \sum_{i=1}^n T_i \log(l^{-1}(\phi_m(X_i)^T \theta)) + (1 - T_i) \log(1 - l^{-1}(\phi_m(X_i)^T \theta)).$$

This is a special case of the proper scoring rule maximization (4) when the rule S is $S_{0,0}$ in the Beta family.

Besides Assumption 1 (strong ignorability), the other technical assumptions in Hirano et al. (2003) are listed below.

Assumption 2. (*Distribution of X*) The support of X is a Cartesian product of compact intervals. The density of X is bounded, and bounded away from 0.

Assumption 3. (*Distribution of $Y(0)$, $Y(1)$*) The second moments of $Y(0)$ and $Y(1)$ exist and $g(X, 0) = E[Y(0)|X]$ and $g(X, 1) = E[Y(1)|X]$ are continuously differentiable.

Assumption 4. (*Propensity score*) The propensity score $p(X) = P(T = 1|X)$ is continuously differentiable of order $s \geq 7d$ where d is the dimension of X , and $p(x)$ is bounded away from 0 and 1.

Assumption 5. (*Sieve estimation*) *The nonparametric sieve logistic regression uses a power series with $m = n^\nu$ for some $1/(4(s/d - 1)) < \nu < 1/9$.*

The most notable assumptions are the compactness of the support of X (Assumption 2) and the smoothness of $p(X)$ (Assumption 4), which are generally required in nonparametric regression, and the strong overlap assumption that $p(X)$ is bounded away from 0 and 1 (Assumption 4), which is necessary to ensure generalized inverse probability weight (9) is bounded. Another important assumption is the rate $m = n^\nu$ as $n \rightarrow \infty$ (Assumption 5).

Theorem 2 below is an extension to the main theorem of Hirano et al. (2003). Compared to the original theorem which always uses the maximum likelihood for any weighted average treatment effect, the scoring rule is now tailored according to the estimand as described in Section 3.

Theorem 2. *Suppose we use the Beta-family of covariate balancing scoring rules defined by equations (2) and (3) with $-1 \leq \alpha, \beta \leq 0$ and the logistic link (6). Under Assumptions 1 to 5, the propensity score weighting estimator $\hat{\tau}_{\alpha, \beta}$ and its normalized version $\hat{\tau}_{\alpha, \beta}^*$ are consistent for $\tau_{\alpha, \beta}$ and $\tau_{\alpha, \beta}^*$. Moreover, they reach the semiparametric efficiency bound for estimating $\tau_{\alpha, \beta}$ and $\tau_{\alpha, \beta}^*$.*

Proof. See Appendix A.2. □

5.2. Implications of Covariate Balance. If there is no efficiency gain in using maximum likelihood, what are the benefits of using a CBSR so the predictors are automatically balanced? One benefit is that the inverse weights w are less volatile, thanks to the observation in Section 3 that CBSR penalizes extreme inverse probability weights. This Section discusses another advantage, namely the bias reduction of $\hat{\tau}$ when $p_\theta(X)$ is misspecified. This is perhaps more important in practice, as Box (1976) once said: “all models are wrong, but some are useful”.

Here we investigate the bias of $\hat{\tau}$ under the global null model. Denote the true outcome regression functions by $g(X, t) = E[Y(t)|X]$, $t = 0, 1$. In the global null model, $g(x, 1) = g(x, 0)$ for all x , so there is no treatment effect whatsoever. By definition (11) and (12), the weighted average treatment effects $\tau = \tau^*$ are always equal to 0.

Suppose the propensity score model is specified by $p_\theta(X)$ and the corresponding weights (9) are $w_\theta(X)$. Let $\tilde{\theta} = \arg \max_\theta S(p_\theta(X), p(X))$, so $p_{\tilde{\theta}}(x)$ is the best approximation of $p(x)$ with respect to the scoring rule S . Furthermore, define

$$\tilde{w}(x) = \frac{e(x)w_{\tilde{\theta}}(x, 1)}{E[e(X)w_{\tilde{\theta}}(X, 1)]} - \frac{(1 - e(x))w_{\tilde{\theta}}(x, 0)}{E[(1 - e(X))w_{\tilde{\theta}}(X, 0)]}.$$

The asymptotic bias of $\hat{\tau}^*$ is given by

$$\text{bias}(\hat{\tau}^*) = E[\hat{\tau}^*] = E[\tilde{w}(X)g(X)].$$

When $p_\theta(X)$ is correctly specified (i.e. $p_{\tilde{\theta}}(x) = p(x)$), by the definition of GIPW (9), $\tilde{w}(x)$ is always zero. Therefore $\hat{\tau}$ is asymptotically unbiased under correctly specified propensity score model. When $p_\theta(X)$ is not correctly specified, the bias of $\hat{\tau}$ heavily depends on the covariate balance under the weight $w_{\tilde{\theta}}(X)$. To see this, notice that the covariate balancing

property (10) can be written as $E[\tilde{w}(X)\phi(X)] = 0$. Therefore, for any $\eta \in \mathbb{R}^m$,

$$\begin{aligned} \text{bias}(\hat{\tau}^*) &= E[\tilde{w}(X)(g(X) - \eta^T \phi(X))] \\ &\leq E|\tilde{w}(X)| \cdot \left(\sup_x |g(x) - \eta^T \phi(x)| \right) \\ &= 2 \sup_x |g(x) - \eta^T \phi(x)|. \end{aligned} \quad (22)$$

The last inequality is true if $\phi(x)$ includes an intercept term, since by (10),

$$E[e(x)w_{\hat{\theta}}(X, 1)] = E[Tw_{\hat{\theta}}(X, 1)] = 1 = E[(1 - T)w_{\hat{\theta}}(X, 0)] = E[(1 - e(X))w_{\hat{\theta}}(X, 1)].$$

Equation (22) leads to the next result:

Theorem 3. *Under Assumption 1 and the global null that $g(x, 0) = g(x, 1) = g(x)$ for all x , the estimator $\hat{\tau}^*$ is asymptotically unbiased if*

- (i) *A covariate balancing scoring rule is used and $\phi(x)$ includes an intercept term, and*
- (ii) *$g(x)$ is in the linear span of $\{\varphi_1(x), \dots, \varphi_m(x)\}$, or more generally $\inf_{\eta} \|g(x) - \eta^T \phi_m(x)\|_{\infty} \rightarrow 0$ as $n, m(n) \rightarrow \infty$.*

The last condition says that $g(x)$ can be uniformly approximated by functions in the linear span of $\phi_1(x), \dots, \phi_m(x)$ as $m \rightarrow \infty$. This holds under very mild assumption of g . For example, if the support of X is compact and $g(x)$ is continuous, the Weierstrass approximation theorem ensures that $g(x)$ can be uniformly approximated by polynomials. Theorem 3 can also be easily extended to the constant treatment effect model $g(x, 1) = g(x, 0) + c$. In this case, $\tau^* = c$ under any weighting and one can verify that the upper bound in (22) still holds.

Finally we compare the results in Theorem 3 and Theorem 2. The main difference is that Theorem 2 uses *propensity score* models with increasing complexity, whereas Theorem 3 assumes uniform approximation for the *outcome regression* function. Since the unbiasedness in Theorem 3 does not presume any assumption on the propensity score, the estimator $\hat{\tau}$ obtained by CBSR is more robust to misspecified or overfitted propensity score model.

5.3. Lagrangian Duality. To understand the fundamental connection between propensity score weighting and empirical calibration in survey sampling (Deville and Särndal, 1992), here we present an alternative way to derive CBSR through Lagrangian duality. First, let's rewrite the score optimization problem (4) by introducing new variables f_i for each observation i :

$$\begin{aligned} &\underset{f, \theta}{\text{maximize}} \quad \frac{1}{n} \sum_{i=1}^n S(l^{-1}(f_i), T_i) \\ &\text{subject to} \quad f_i = \theta^T \phi(X_i), \quad i = 1, \dots, n. \end{aligned} \quad (23)$$

Let the Lagrangian multiplier associated with the i -th constraint be $(2T_i - 1)w_i/n$. The notation w indicates inverse probability weights in the last section. The reason of this abuse of notation will become clear in a moment. The Lagrangian of (23) is given by

$$\text{Lag}(f, \theta; w) = \frac{1}{n} \sum_{i=1}^n S(l^{-1}(f_i), T_i) + (2T_i - 1)w_i [f_i - \theta^T \phi(X_i)].$$

By setting the partial derivatives of the Lagrangian equal to 0, we obtain

$$\frac{\partial \text{Lag}}{\partial \theta_k} = \frac{1}{n} \sum_{i=1}^n (2T_i - 1) w_i \phi_k(X_i) = 0, \quad k = 1, \dots, m. \quad (24)$$

$$\frac{\partial \text{Lag}}{\partial f_i} = \frac{1}{n} \left(\frac{\partial S(l^{-1}(f_i), T_i)}{\partial f_i} + (2T_i - 1) w_i \right) = 0, \quad i = 1, \dots, n, \quad (25)$$

Equation (24) is the same as (16), meaning the optimal dual variables w balance the predictors ϕ_1, \dots, ϕ_m . Equation (25) determines w from f . By using (2) and the logistic link (6), it turns out that $w_i = w(X_i, T_i)$ is exactly the GIPW weights defined in (9). In conclusion, the weights w are the dual variables of the score optimization problem (4) and are required to balance the predictors ϕ .

The benefit of this derivation is that we can write down the Lagrangian dual problem of (23). In general, there is no explicit form for $-1 < \alpha, \beta < 0$ because it is difficult to invert (9), but in the particularly interesting cases $\alpha = 0, \beta = -1$ (corresponding to ATT) and $\alpha = -1, \beta = -1$ (corresponding to ATE), the dual problems are algebraically tractable. When $\alpha = 0, \beta = -1$, the treated units are weighted by 1 and the control units are weighted by $\hat{p}/(1 - \hat{p})$. In this case, the Lagrangian dual optimization problem is given by

$$\begin{aligned} & \underset{w \geq 0}{\text{minimize}} && \sum_{i:T_i=0} w_i \log w_i - w_i \\ & \text{subject to} && \sum_{i:T_i=0} w_i \phi_k(X_i) = \sum_{j:T_j=1} \phi_k(X_j), \quad k = 1, \dots, m. \end{aligned} \quad (26)$$

In most cases an intercept term is included in the GLM, so the constraints in (26) imply that $\sum_{T_i=0} w_i$ is equal to the number of treated units (a fixed value). Therefore the dual optimization problem is equivalent to the following maximum entropy problem

$$\begin{aligned} & \underset{w \geq 0}{\text{minimize}} && \sum_{i:T_i=0} w_i \log w_i \\ & \text{subject to} && \sum_{i:T_i=0} w_i \phi_k(X_i) = \sum_{j:T_j=1} \phi_k(X_j), \quad k = 1, \dots, m. \end{aligned} \quad (27)$$

When $\alpha = \beta = -1$, the inverse probability weights are always greater than 1. It turns out that the Lagrangian dual problem in this case is given by

$$\begin{aligned} & \text{minimize} && \sum_{i=1}^n (w_i - 1) \log(w_i - 1) - w_i \\ & \text{subject to} && \sum_{i:T_i=0} w_i \phi_k(X_i) = \sum_{j:T_j=1} w_j \phi_k(X_j), \quad k = 1, \dots, m. \\ & && w_i \geq 1, \quad i = 1, \dots, n. \end{aligned} \quad (28)$$

The objective functions in (26) and (28) encourage w to be close to uniform. They belong to a general distance measure $\sum_{i=1}^n D(w_i, v_i)$ in Deville and Särndal (1992), where $D(w, v)$ is a continuously differentiable and strongly convex function in w and achieves its minimum at (the limit) $w = v$. When the estimand is ATT (or ATE), the target v is equal to 1 (or 2). The average treatment effect estimators of this kind are often called “calibration estimators” in the survey sampling literature, because the weighted sample averages are empirically calibrated to some known unweighted population averages.

The maximum entropy problem (27) appeared first in Hainmueller (2011) to estimate ATT and is called “Entropy Balancing”. Zhao and Percival (2015) used the primal-dual connection described above to show Entropy Balancing is doubly robust, which is stronger than Theorems 2 and 3. Unfortunately, the double robustness only holds when the estimand is ATT. This section generalizes the primal-dual connection to other weighted average treatment effects. Chan et al. (2015) studied the calibration estimators with the general distance D and showed the estimator $\hat{\tau}$ is globally semiparametric efficient. When the estimand is ATE, Chan et al. (2015) require the weighted sums of ϕ_k in (28) to be calibrated to $\sum_{i=1}^n \phi_k(X_i)/n$, too. It is shown earlier in Section 5.1 that this extra calibration is not necessary for semiparametric efficiency.

In an extension to Entropy Balancing, Hazlett (2013) proposed to empirically balance kernel representers instead of ordinary predictors. This corresponds to unregularized ($\lambda = 0$) RKHS regression introduced in Section 4.3. The unregularized problem is unfeasible if the RKHS is rich, so Hazlett (2013) tweaked the objective in order to find a usable solution.

5.4. Bias-variance tradeoff. The results in Sections 5.2 and 5.3 allow us to study the fundamental bias-variance in selecting a propensity score model. Consider the regularized regression approach introduced in Section 4.2. By the Karush-Kuhn-Tucker conditions of the regularized score maximization problem (4), the solution $\hat{\theta}_\lambda$ satisfies (for any $a \geq 1$ and $\lambda \geq 0$)

$$\left| \sum_{T_i=1} w_{\hat{\theta}_\lambda}(X_i, T_i) \phi_k(X_i) - \sum_{T_i=0} w_{\hat{\theta}_\lambda}(X_i, T_i) \phi_k(X_i) \right| \leq \lambda \cdot |(\hat{\theta}_\lambda)_k|^{a-1}, k = 1, \dots, m. \quad (29)$$

The equality in (29) holds if $(\hat{\theta}_\lambda)_k \neq 0$, which is true unless $a = 1$ and λ is large. This suggests that in general the predictors $\phi(x)$ are not exactly balanced when $\lambda > 0$.

Following Section 5.2, we assume the null model $E[Y(1)|X] = E[Y(0)|X] = g(x)$ to study how covariate imbalance affects the bias of $\hat{\tau} = \hat{\tau}_\lambda$. Moreover, let's assume the outcome regression function is in the linear span of the predictors, i.e. $g(x) = g_\eta(x) = \sum_{j=1}^m \eta_j \phi_j(x)$. The finite sample bias of $\hat{\tau}_\lambda$ under the null is given by

$$\begin{aligned} \text{bias}_\eta(\hat{\tau}_\lambda) &= \left| \sum_{T_i=1} w_{\hat{\theta}_\lambda}(X_i, T_i) g_\eta(X_i) - \sum_{T_i=0} w_{\hat{\theta}_\lambda}(X_i, T_i) g_\eta(X_i) \right| \\ &= \left| \sum_{j=1}^m \eta_j \left(\sum_{T_i=1} w_{\hat{\theta}_\lambda}(X_i, T_i) \phi_j(X_i) - \sum_{T_i=0} w_{\hat{\theta}_\lambda}(X_i, T_i) \phi_j(X_i) \right) \right| \\ &\leq \lambda \left| \sum_{k=1}^m \eta_k |(\hat{\theta}_\lambda)_k|^{a-1} \right| \leq \lambda \|\eta\|_a \|\hat{\theta}_\lambda\|_a^{a-1}. \end{aligned}$$

The last inequality is due to Hölder's inequality and is tight. Hence we have

$$\max_{\eta} \frac{\text{bias}_\eta(\hat{\tau}_\lambda)}{\|\eta\|_a} = \lambda \|\hat{\theta}_\lambda\|_a^{a-1}. \quad (30)$$

The next proposition says that the right hand side of equation (30) is decreasing as the degree of regularization λ becomes smaller. This is consistent with our intuition that the more we regularize the propensity score model, the more bias we get.

Proposition 3. *Given a strictly proper scoring rule S and a link function l such that $S(l^{-1}(f), t)$ is strongly concave and second order differentiable in $f \in \mathbb{R}$ for $t = 0, 1$, let $\hat{\theta}_\lambda$ be the solution to (18) and (19) for a given $a \geq 1$. Then $\lambda \|\hat{\theta}_\lambda\|_a^{a-1}$ is a strictly increasing function of $\lambda > 0$.*

Proof. See Appendix A.3. □

The bias-variance tradeoff is more apparent in the Lagrangian dual problem of (18). Consider the case that the estimand is ATT and the corresponding scoring rule is $\alpha = 0$ and $\beta = -1$. When $\phi_1(x) = 1$ and $J(\theta) = \sum_{k=2}^m |\theta_k|^a / a$ so we do not penalize the intercept, the dual problem of (18) is given by

$$\begin{aligned} & \underset{w \geq 0}{\text{minimize}} && \sum_{i: T_i=0} w_i \log w_i \\ & \text{subject to} && b_k = \sum_{i: T_i=0} w_i \phi_k(X_i) - \sum_{j: T_j=1} \phi_k(X_j), \quad k = 1, \dots, m. \\ & && b_1 = 0, \quad \|b\|_{a/(a-1)} \leq r(\lambda), \end{aligned} \tag{31}$$

where $r(\lambda)$ is some increasing function of λ . In (31), the objective function measures the closeness between w and the uniform weights and the constraints bound the covariate imbalance with respect to the functions ϕ . They are related, respectively, to the variance and bias of the estimator $\hat{\tau}$. When $\lambda \rightarrow 0$, the solution of (31) converges to the weights w that minimizes the $a/(a-1)$ -norm of covariate imbalance. The limit of $r(\lambda)$ when $\lambda \rightarrow 0$ can be 0 or some positive value, depending on if the unregularized score maximization problem (4) is feasible or not. When $\lambda \rightarrow \infty$, the solution of (31) converges to uniform weights (i.e. no adjustment at all) whose estimator $\hat{\tau}$ has smallest variance. Similar arguments hold if we change the estimand (e.g. to ATE) and the scoring rule accordingly.

The kernel method introduced in Section 4.3 is a special case of the regularized regression with potentially infinite-dimensional predictors. For RKHS regressions, the maximum bias (30) under the null model is given by

$$\max_{g \in \mathcal{H}_K} \frac{\text{bias}_g(\hat{\tau}_\lambda)}{\|g\|_{\mathcal{H}_K}} = \lambda \|f\|_{\mathcal{H}_K}.$$

Therefore, the bias of $\hat{\tau}$ is controlled for a rich class of outcome regression functions.

6. NUMERICAL EXAMPLES

We use two examples (one simulation and one real data) to demonstrate the effectiveness of CBSR and the proposed adaptive methods.

6.1. A simulation example. This example due to Kang and Schafer (2007) is also used to generate Figure 1 in the Introduction. The artificial dataset consists of i.i.d. random variables (X_i, Z_i, T_i, Y_i) , $i = 1, \dots, n$, where X_i , Y_i and T_i are always observed and Z_i is never observed. To generate this data set, X_i is a 4-dimensional vector distributed as $N(0, I_4)$; Z_i is computed

by first applying the following transformation:

$$\begin{aligned} Z_{i1} &= \exp(X_{i1}/2), \\ Z_{i2} &= X_{i2}/(1 + \exp(X_{i1})) + 10, \\ Z_{i3} &= (X_{i1}X_{i3} + 0.6)^3, \\ Z_{i4} &= (X_{i2} + X_{i4} + 20)^2, \end{aligned}$$

and then normalizing individual variables of Z to have sample mean 0 and variance 1.

There are in total four settings in this example. In the first setting (top-left panel in Figure 3), Y_i is generated by $Y_i = g(X, T_i)$ without any additional noise, where

$$g(X, 0) = 210 + 27.4X_{i1} + 13.7X_{i2} + 13.7X_{i3} + 13.7X_{i4},$$

and $g(X, 1)$ is either equal to $g(X, 0)$ (column “zero” in Figure 3) or $g(X, 0) + 10$ (column “constant” in Figure 3). The true propensity scores are generated by the logistic model $p(X_i) = l^{-1}(f(X_i))$, $f(X_i) = -X_{i1} + 0.5X_{i2} - 0.25X_{i3} - 0.1X_{i4}$. In this setting, both Y and T can be correctly modeled by (generalized) linear model of the observed covariates X . In the other settings, at least one of the propensity score model and the outcome regression model is non-linear in X . In order to achieve this, the data generating process described above is altered such that Y or T (or both) is linear in the unobserved Z instead of the observed X , though the parameters are kept the same. In these three scenarios, at least one of the two functions f and g are nonlinear in X .

For each setting, 200 replicas of dataset of size $n = 200$ are drawn. The logistic link function is always used and different scoring rules in the Beta-family (3) are applied. The predictor vector ϕ used is $\phi(X) = (X_1, X_1^2, X_2, X_2^2, X_3, X_3^2, X_4, X_4^2)$. After an estimated propensity score model is obtained, we use the normalized IPW estimator $\hat{\tau}_{-1, -1}^*$ to estimate ATE and $\hat{\tau}_{0, -1}^*$ to estimate ATT. The covariate imbalance with respect to ϕ is shown earlier in Figure 1.

Figure 3 shows the boxplots of these estimates under different settings. It is clear that the covariate balancing scoring rules (CBSR) generate much more stable estimates than the Bernoulli likelihood (MLE). Furthermore, in the two left panels the true logit f is linear in X so the propensity score model is correctly specified. In the two top panels the true outcome regression function g_0 is linear in X so the unbiasedness is guaranteed by Theorem 3. As expected, the weighting estimators given by CBSR are unbiased across these three panels (besides the bottom-right panel). If instead the Bernoulli likelihood criterion is used to estimate the propensity score model, the weighting estimator is biased when f is non-linear in X even if g is linear in X (top-right panel). Even if f is linear in X so the propensity score model is correctly specified, the CBSR estimators have much smaller variance than MLE. Lastly, in the bottom-right panel where both f and g are non-linear, CBSR still has smaller bias and variance.

Next we test the adaptive strategies described in Section 4. Here we consider three adaptive strategies—forward stepwise regression and two reproducing kernel Hilbert space (RKHS) regressions. In the forward stepwise regression, we use all the two-way interactions of degree-two polynomials (in total 32 predictors) to allow sophisticated propensity score models. In the two RKHS regressions, we use the Gaussian kernel

$$K(x, x') = \exp(-\gamma\|x - x'\|^2), \quad x, x' \in \mathbb{R}^4,$$

with γ equal to 0.2 and 0.5. After fitting a path of propensity score models (indexed by step for forward stepwise and regularization parameter λ for RKHS), for each strategy we

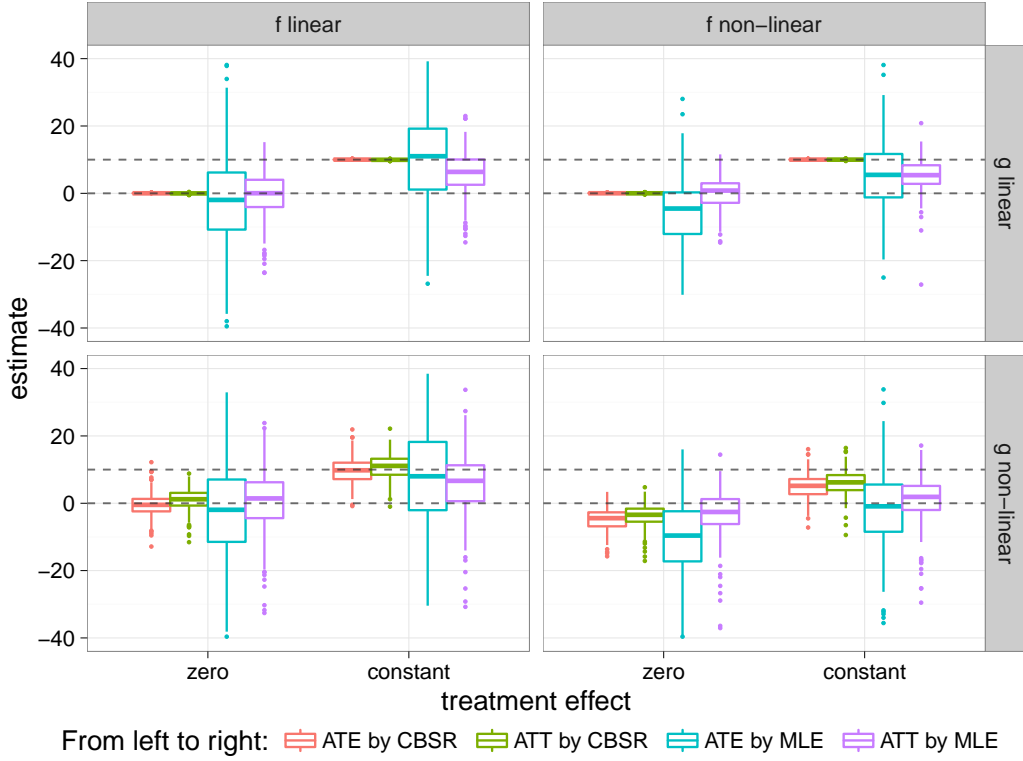


FIGURE 3. Estimate of average treatment effects (ATE or ATT) using different scoring rules under the four settings. The four boxes in each group with different colors correspond to the Beta scoring rule (3) with 1. $\alpha = \beta = -1$; 2. $\alpha = 0, \beta = -1$; 3. $\alpha = \beta = 0$; 4. $\alpha = \beta = 0$ (Bernoulli likelihood). In the first and third boxes, the inverse probability weights corresponding to ATE ($\alpha = \beta = -1$ in equation (15)) is used. In the second and fourth boxes, the inverse probability weights corresponding to ATT ($\alpha = 0, \beta = -1$) is used. The gray dashed lines correspond to the true treatment effect: 0 for group “zero” and 10 for group “constant”.

choose the model that has the smallest number of significant two-sample t -tests as described in Section 4.4. Finally, the standard errors and confidence intervals by assuming homoscedasticity in (21) and using a pilot outcome regression (predictors are the observed X). Since a fairly sophisticated propensity score model can be fitted, we use $n = 1000$ samples and set $g(X, 1) = g(X, 0)$ to test all the methods.

Table 2 shows the performance of the six different combinations of loss function and adaptive strategy in the four simulation settings. CBSR clearly outperforms Bernoulli likelihood. In almost all scenarios and no matter what adaptive strategy is used, the root mean squared error (RMSE) of CBSR is less than half of the RMSE of Bernoulli likelihood. The confidence intervals obtained by using Bernoulli likelihood also perform poorly. In many scenarios the actual coverage is less than 50%, whereas the nominal coverage is 95%. CBSR’s confidence intervals have close to or over the nominal 95% coverage in almost all scenarios.

The two adaptive strategies (forward stepwise and RKHS) perform similarly. When using CBSR as the loss function, forward stepwise seems to have slightly smaller RMSE, but kernel methods can have the better coverage in some scenarios. In practice, the user may want to choose an adaptive strategy that is most convenient for the target application. The strength and weakness of these methods are discussed in Section 4.

6.2. A Real Data Example. This Section studies the National Supported Work (NSW) Demonstration which was previously analyzed by LaLonde (1986), Dehejia and Wahba (1999), Smith and Todd (2005) and many other authors. The NSW Demonstration was a federally and privately funded program implemented in the 1970s, which program provided transitional and subsidized work experience for a period of 6–18 months to individuals who had faced economic and social problems prior to the enrollment in the program. The pre-treatment covariates include earnings, education, age, ethnicity, and marital status, and the outcome of interest in LaLonde (1986) is the post-intervention earnings in 1978. We use the experimental subsample taken by Dehejia and Wahba (1999) to demonstrate our methods, which include 185 treated and 260 control observations that joined the program early enough for the retrospective earnings information in the year 1974. To evaluate the non-experimental methods, we use the Current Population Survey (CPS) data extracted by LaLonde (1986) as the control group, which contain 15992 observations. The reader is referred to the previous articles listed in this paragraph for more detailed information on this dataset.

The observational methods are evaluated in two scenarios:

1. Compare the experimental treated group with the non-experimental control group. The average treatment effect estimators can be compared with the experimental benchmark, which is 1794.3 (standard error 632.9) by a linear regression of the earnings in 1978 on the treatment assignment.
2. Compare the experimental control group with the non-experimental control group. Since both groups did not receive treatment, the treatment effect is always zero (the null case in Section 5.2).

Since the non-experimental control group is very large and has very different covariate distribution to the experimental group, we only consider the average treatment effect on the treated (ATT) in this example. Using Table 1, the CBSR rule in this case is $S_{0,-1}$.

First, we apply the forward stepwise regression in Algorithm 1. To generate predictors $\phi(X)$, we use all the discrete covariates (race, married, no degree, no earning in 1974, no earning in 1975), all the continuous covariates (age, year of education, earning in 1974, earning in 1975) and their squares, and the first-order interactions of all these variables. This results in a 94-dimensional vector ϕ of predictors. In each scenario, two scoring rules, the Bernoulli likelihood $S_{0,0}$ and the covariate balancing scoring rule $S_{0,-1}$, are used, and the covariate imbalance and the estimator $\hat{\tau}$ are plotted along the stepwise path.

The results of the forward stepwise regressions can be found in Figure 4. Compared to the Bernoulli likelihood, CBSR requires a stronger condition for the existence of the solution (Zhao and Percival, 2015), so in both scenarios CBSR stops early (71 steps in scenario 1 and 23 steps in scenario 2). Nevertheless, CBSR is much better at reducing covariate imbalance as shown in Figures 4a and 4b. In fact, at least 10 predictors have standardized difference greater than 20% across the entire path when the Bernoulli likelihood is used, while some authors have suggested that a standardized difference above 10% can be deemed substantial (Normand et al., 2001, Austin and Stuart, 2015). When a two-sample t-test is used to compare the mean of the predictors, less than 20% of the 94 tests are insignificant with the

f (true PS)	g (true OR)	estimand	loss	strategy	bias	RMSE	coverage
linear	linear	ATE	Bernoulli	forward stepwise	-1.27	2.43	50.5
				kernel (0.2)	-2.77	3.18	22.5
				kernel (0.5)	-3.05	3.45	17.5
			CBSR	forward stepwise	-0.24	0.98	90.0
				kernel (0.2)	-0.50	1.16	90.5
				kernel (0.5)	-1.41	1.77	63.0
		ATT	Bernoulli	forward stepwise	-2.92	6.28	55.5
				kernel (0.2)	-7.30	11.08	29.0
				kernel (0.5)	-2.70	3.34	27.5
			CBSR	forward stepwise	-0.24	1.19	91.5
				kernel (0.2)	-1.09	2.78	90.5
				kernel (0.5)	-1.91	2.49	63.5
	nonlinear	ATE	Bernoulli	forward stepwise	-1.05	2.17	82.5
				kernel (0.2)	-1.90	2.52	74.0
				kernel (0.5)	-2.13	2.77	68.0
			CBSR	forward stepwise	-0.46	1.17	99.5
				kernel (0.2)	-0.46	1.17	100.0
				kernel (0.5)	-1.10	1.62	96.0
		ATT	Bernoulli	forward stepwise	-0.93	3.87	87.5
				kernel (0.2)	-3.16	6.64	65.5
				kernel (0.5)	-0.05	1.82	94.0
			CBSR	forward stepwise	-0.49	1.37	99.0
				kernel (0.2)	-0.27	2.05	99.0
				kernel (0.5)	-0.59	1.83	99.5
nonlinear	linear	ATE	Bernoulli	forward stepwise	-1.55	2.07	45.5
				kernel (0.2)	-2.28	2.75	31.0
				kernel (0.5)	-2.54	2.97	24.5
			CBSR	forward stepwise	-0.27	1.02	86.5
				kernel (0.2)	-0.40	1.10	92.5
				kernel (0.5)	-1.19	1.61	64.5
		ATT	Bernoulli	forward stepwise	-0.45	1.94	78.5
				kernel (0.2)	-0.88	2.43	64.0
				kernel (0.5)	-0.92	1.84	62.0
			CBSR	forward stepwise	-0.13	1.14	89.0
				kernel (0.2)	-0.36	1.46	93.5
				kernel (0.5)	-0.83	1.61	82.5
	nonlinear	ATE	Bernoulli	forward stepwise	-2.25	2.75	64.5
				kernel (0.2)	-2.90	3.37	51.0
				kernel (0.5)	-3.29	3.69	41.5
			CBSR	forward stepwise	-0.61	1.01	100.0
				kernel (0.2)	-0.73	1.26	100.0
				kernel (0.5)	-1.88	2.19	86.5
		ATT	Bernoulli	forward stepwise	-0.12	1.82	96.5
				kernel (0.2)	-0.02	2.34	96.0
				kernel (0.5)	-0.37	1.64	95.5
			CBSR	forward stepwise	-0.39	1.12	100.0
				kernel (0.2)	-0.35	1.46	100.0
				kernel (0.5)	-1.01	1.74	99.5

TABLE 2. Performance of different loss functions combined with adaptive strategies—forward stepwise and kernel method (Gaussian kernel with bandwidth parameter 0.2 and 0.5). In each case, the propensity score model is selected to minimize the number of significant covariate imbalance tests. Compared to the Bernoulli likelihood, maximizing the covariate balancing scoring rule (CBSR) reduces the root mean square error (RMSE) by more than a half for most settings. CBSR’s confidence intervals also have the superior coverage (nominal level is 95%).

Bernoulli likelihood, also implying insufficient covariate balance. On the contrary, CBSR successfully balances most predictors in both scenarios.

Figures 4c and 4d show the estimate of ATT along the path. Interesting, by just including the first predictor most of the bias of estimating ATT is corrected. Both scoring rules give similar estimates and are consistent with the experimental benchmarks. However, as discussed above, the weights generated by maximizing the Bernoulli likelihood are unacceptable to many applied researchers. Switching to CBSR solves this problem, though the ATT estimates are not very different in this particular example. Additionally, when using CBSR the standard error of $\hat{\tau}$ is smaller. This can be understood from the remark in Section 3.2 that CBSR tries to avoid large weights.

Next, we apply the kernel method in Section 4.3 and the results are presented in Figure 5. We use the Gaussian kernel $K(x, x') = \exp(-\sigma\|x - x'\|^2)$ with $\sigma = 0.15$ and $x = (\text{black}, \text{hispanic}, \text{no degree}, \text{married}, \text{age}/5, \text{education}/3, \text{re74}/4000, \text{re75}/4000)$. The first four entries in x are indicator variables, and **re74** (**re75**) stands for the annual earning of the person in the year 1974 (1975). Because the kernel matrix is a large $n \times n$ matrix, we use the subsample CPS2 extracted by Dehejia and Wahba (1999) that contains $n = 2369$ non-experimental controls. Overall, these two plots are similar to those for the forward stepwise regressions. Notice that the confidence intervals of ATT are wider when using the kernel method. This loss of efficiency is compensated by the improved robustness, as the propensity score weights approximately balance infinite many covariate functions (Section 4.3).

7. CONCLUSIONS

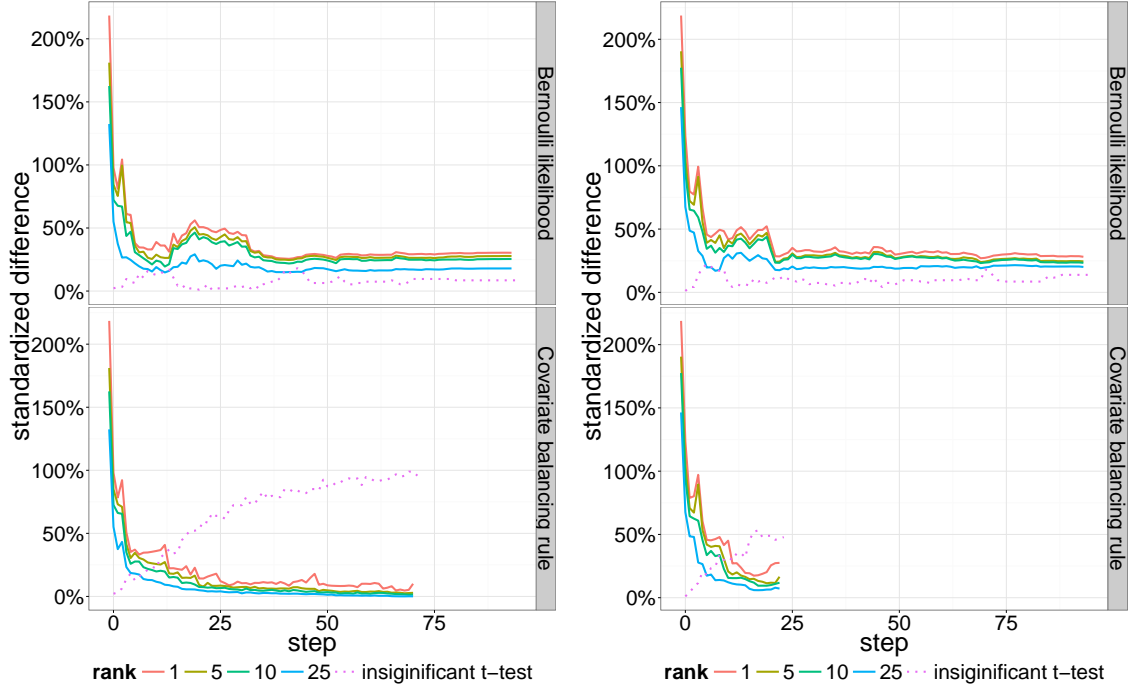
We have derived a new family of scoring rules to estimate the propensity score. If the covariate balancing scoring rule (CBSR) is used to fit a generalized linear model, exact sample balance can be achieved. The scoring rule can also be easily used in machine learning algorithms to optimize the propensity score model. Our numerical examples show that CBSR's empirical performance dominates the routinely used Bernoulli likelihood.

APPENDIX A. TECHNICAL PROOFS

A.1. Proof of Proposition 2. The same result can be found in Buja et al. (2005, Section 15). For completeness we give a direct proof here. Denote $p = l^{-1}(f) \in (0, 1)$ and notice that $df/dp = (l^{-1})'(f) = p(1 - p)$. By (2), we have

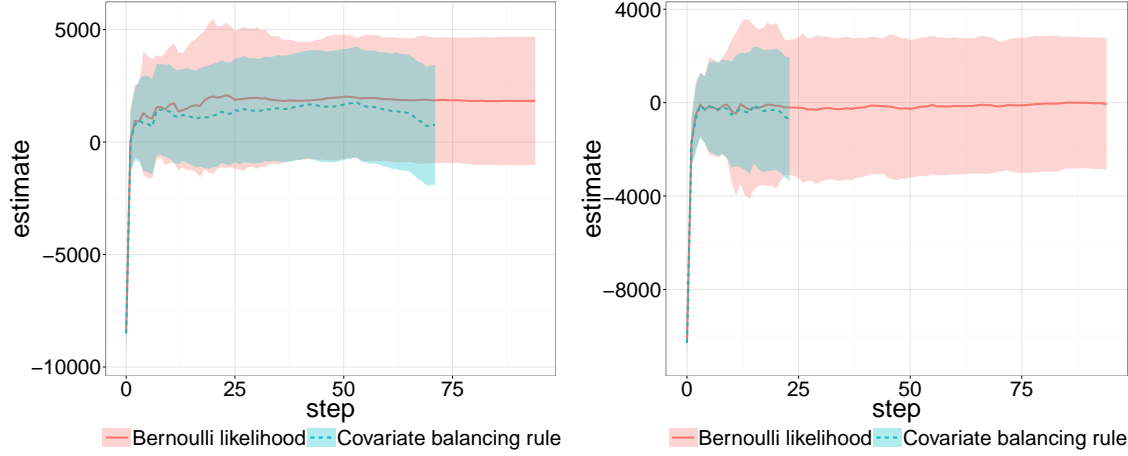
$$\begin{aligned} \frac{d}{df}S(l^{-1}(f), 1) &= (1 - p)G''(p)(l^{-1})'(f) = p^\alpha(1 - p)^{\beta+1}, \\ \frac{d}{df}S(l^{-1}(f), 0) &= -pG''(p)(l^{-1})'(f) = -p^{\alpha+1}(1 - p)^\beta, \text{ and} \\ \frac{d^2}{df^2}S(l^{-1}(f), 1) &= \alpha p^\alpha(1 - p)^{\beta+2} - (\beta + 1)p^{\alpha+1}(1 - p)^{\beta+1}, \\ \frac{d^2}{df^2}S(l^{-1}(f), 0) &= -(\alpha + 1)p^{\alpha+1}(1 - p)^{\beta+1} + \beta p^{\alpha+2}(1 - p)^\beta. \end{aligned}$$

The conclusions immediate follow by letting the second order derivatives be less than or equal to 0.



(A) Covariate imbalance: experimental treatment vs. non-experimental control.

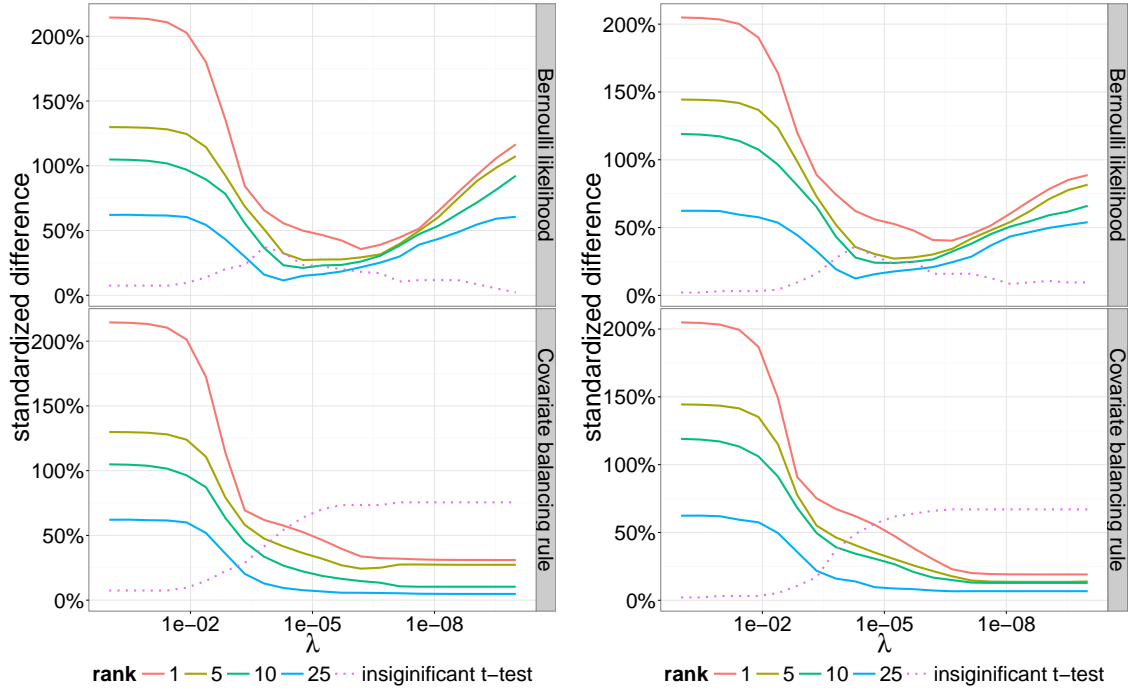
(B) Covariate imbalance: experimental control vs. non-experimental control.



(C) Estimate of ATT: experimental treatment vs. non-experimental control.

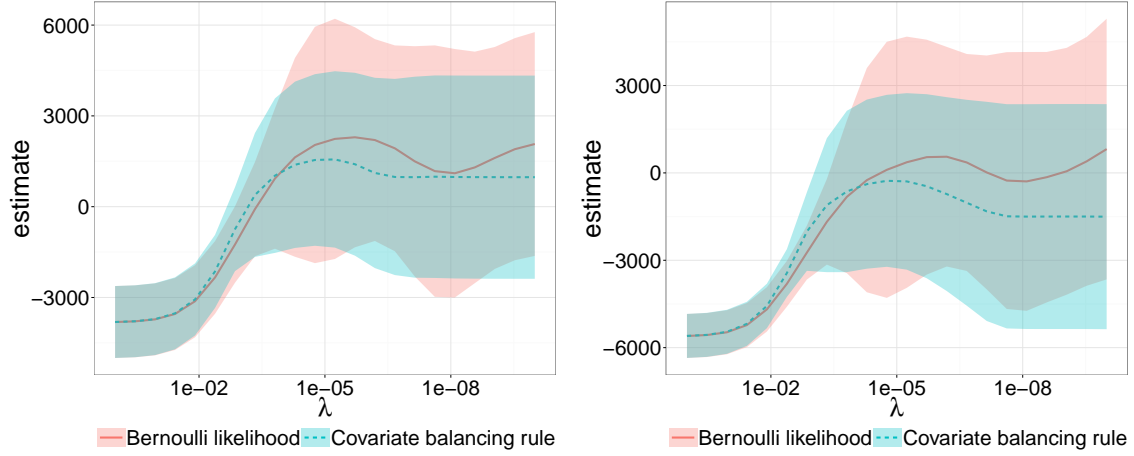
(D) Estimate of ATT: experimental control vs. non-experimental control.

FIGURE 4. Forward stepwise regressions for the LaLonde (1986) dataset. Top panels: Covariate imbalance is in terms of standardized difference. The curves in the plot are the 1st, 5th, 10th, and 25th largest standardized difference among the 94 predictors (solid lines), and the percentage of insignificant two-sample t-tests comparing the mean of each predictor in the treatment and the weighted control (dotted line). Bottom panels: estimated ATT with 95% confidence interval.



(A) Covariate imbalance: experimental treatment vs. non-experimental control.

(B) Covariate imbalance: experimental control vs. non-experimental control.



(C) Estimate of ATT: experimental treatment vs. non-experimental control.

(D) Estimate of ATT: experimental control vs. non-experimental control.

FIGURE 5. Reproducing kernel method for the LaLonde (1986) dataset. Top panels: Covariate imbalance is in terms of standardized difference. The curves in the plot are the 1st, 5th, 10th, and 25th largest standardized difference among the 94 predictors (solid lines), and the percentage of insignificant two-sample t-tests comparing the mean of each predictor in the treatment and the weighted control (dotted line). Bottom panels: estimated ATT with 95% confidence interval.

A.2. Proof of Theorem 2. The proof is a simple modification of the proof in Hirano et al. (2003). In fact, Hirano et al. (2003) only proved the convergence of the estimated propensity score up to certain order. This essentially suggests that the semiparametric efficiency of $\hat{\tau}$ does not heavily depend on the accuracy of the sieve logistic regression.

To be more specific, only three properties of the maximum likelihood rule $S = S_{0,0}$ are used in Hirano et al. (2003, Lemmas 1,2):

1. $\tilde{\theta} = \arg \max_{\theta} S(p_{\theta}, p_{\tilde{\theta}})$ (line 5, page 19), this is exactly the definition of a strictly proper scoring rule (1);
2. The Fisher information matrix

$$\frac{\partial^2}{\partial \theta \partial \theta^T} S(p_{\theta}, p_{\tilde{\theta}}) = E_{\tilde{\theta}} \left\{ \left[\frac{d^2}{df^2} S(l^{-1}(f), T) \Big|_{f=\phi(X)^T \theta} \right] \phi(X) \phi(X)^T \right\}$$

has all eigenvalues uniformly bounded away from 0 for all θ and $\tilde{\theta}$ in a compact set in \mathbb{R}^m , where the expectation on the right hand side is taken over X and $T|X \sim p_{\tilde{\theta}}$.

3. As $m \rightarrow \infty$, with probability tending to 1 the observed Fisher information matrix

$$\frac{\partial^2}{\partial \theta \partial \theta^T} \frac{1}{n} \sum_{i=1}^n S(p_{\theta}(X_i), T_i) = \frac{1}{n} \sum_{i=1}^n \left[\frac{d^2}{df^2} S(l^{-1}(f), T_i) \Big|_{f=\phi(X_i)^T \theta} \right] \phi(X_i) \phi(X_i)^T$$

has all eigenvalues uniformly bounded away from 0 for all θ in a compact set of \mathbb{R}^m (line 7–9, page 21).

Because the approximating functions ϕ are obtained through orthogonalizing the power series, we have $E[\phi(X)\phi(X)^T] = I_m$ and one can show its finite sample version has eigenvalues bounded away from 0 with probability going to 1 as $n \rightarrow \infty$. Therefore a sufficient condition for the second and third properties above is that $S(l^{-1}(f), t)$ is strongly concave for $t = 0, 1$. In Proposition 2 we have already proven the strong concavity for all $-1 \leq \alpha, \beta \leq 1$ except for $\alpha = -1, \beta = 0$ and $\alpha = 0, \beta = -1$. In these two boundary cases, among $S(l^{-1}(f), 0)$ and $S(l^{-1}(f), 1)$ one score function is strongly concave and the other score function is linear in f . One can still prove the second and third properties by using Assumption 4 that the propensity score is bounded away from 0 and 1.

A.3. Proof of Proposition 3. The conclusion is trivial for $a = 1$. Denote

$$h(f, t) = \frac{d}{df} S(l^{-1}(f), t) \text{ and } h'(f, t) = \frac{d}{df} h(f, t), \quad t = 0, 1.$$

Because $S(l^{-1}(f), t)$ is concave in f , we have $h'(f, t) < 0$ for all f . The first-order optimality condition of (18) is given by

$$\frac{1}{n} \sum_{i=1}^n h(\hat{\theta}_{\lambda}^T \phi(X_i), T_i) \phi_k(X_i) + \lambda |(\hat{\theta}_{\lambda})_k|^{a-1} \text{sign}((\hat{\theta}_{\lambda})_k) = 0, \quad k = 1, \dots, m.$$

Let $\nabla \hat{\theta}_{\lambda}$ be the gradient of $\hat{\theta}_{\lambda}$ with respect to λ . By taking derivative of the identity above, we get

$$\left[\frac{1}{n} \sum_{i=1}^n h'(\hat{\theta}_{\lambda}^T \phi_i, T_i) \phi_i \phi_i^T + \lambda(a-1) \text{diag}(|\hat{\theta}_{\lambda}|^{a-2}) \right] \nabla \hat{\theta}_{\lambda} = -|\hat{\theta}_{\lambda}|^{a-1} \text{sign}(\hat{\theta}_{\lambda}),$$

where we used the abbreviation $\phi_i = \phi(X_i)$ and $\theta^a = (\theta_1^a, \dots, \theta_m^a)$. For brevity, let's denote

$$H = \frac{1}{n} \sum_{i=1}^n h'(\hat{\theta}_\lambda^T \phi_i, T_i) \phi_i \phi_i^T \prec 0 \text{ and } G = \lambda(a-1) \text{diag}(|\hat{\theta}_\lambda|^{a-2}).$$

For $a > 1$, the result is proven by showing the derivative of $\lambda \|\hat{\theta}_\lambda\|_a^{a-1}$ is greater than 0.

$$\begin{aligned} \frac{d}{d\lambda} \left(\lambda \|\hat{\theta}_\lambda\|_a^{a-1} \right) &= \|\hat{\theta}_\lambda\|_a^{a-1} + \lambda \frac{d}{d\lambda} \left[\sum_{j=1}^m \left| (\hat{\theta}_\lambda)_j \right|^a \right]^{(a-1)/a} \\ &= \|\hat{\theta}_\lambda\|_a^{a-1} + \lambda(a-1) \|\hat{\theta}_\lambda\|_a^{-1} \sum_{j=1}^m \left| (\hat{\theta}_\lambda)_j \right|^{a-1} (\nabla \hat{\theta}_\lambda)_j \text{sign}((\hat{\theta}_\lambda)_j) \\ &= \|\hat{\theta}_\lambda\|_a^{a-1} - \lambda(a-1) \|\hat{\theta}_\lambda\|_a^{-1} (|\hat{\theta}_\lambda|^{a-1})^T (H + G)^{-1} |\hat{\theta}_\lambda|^{a-1} \\ &> \|\hat{\theta}_\lambda\|_a^{a-1} - \lambda(a-1) \|\hat{\theta}_\lambda\|_a^{-1} (|\hat{\theta}_\lambda|^{a-1})^T G^{-1} |\hat{\theta}_\lambda|^{a-1} \\ &= 0. \end{aligned}$$

REFERENCES

- Abadie, A. and G. W. Imbens (2006). Large sample properties of matching estimators for average treatment effects. *Econometrica* 74(1), 235–267.
- Austin, P. C. and E. A. Stuart (2015). Moving towards best practice when using inverse probability of treatment weighting (iptw) using the propensity score to estimate causal treatment effects in observational studies. *Statistics in Medicine* 34(28), 3661–3679.
- Box, G. E. (1976). Science and statistics. *Journal of the American Statistical Association* 71(356), 791–799.
- Buja, A., W. Stuetzle, and Y. Shen (2005). Loss functions for binary class probability estimation and classification: Structure and applications. *Working draft*.
- Caliendo, M. and S. Kopeinig (2008). Some practical guidance for the implementation of propensity score matching. *Journal of Economic Surveys* 22(1), 31–72.
- Chan, K. C. G., S. C. P. Yam, and Z. Zhang (2015). Globally efficient nonparametric inference of average treatment effects by empirical balancing calibration weighting. *Journal of Royal Statistical Society, Series B (Methodology)* to appear.
- Crump, R., V. J. Hotz, G. Imbens, and O. Mitnik (2006). Moving the goalposts: Addressing limited overlap in the estimation of average treatment effects by changing the estimand. Technical Report 330, National Bureau of Economic Research.
- Dehejia, R. H. and S. Wahba (1999). Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs. *Journal of the American Statistical Association* 94(448), 1053–1062.
- Deville, J.-C. and C.-E. Särndal (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association* 87(418), 376–382.
- Gneiting, T. and A. E. Raftery (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association* 102(477), 359–378.
- Hainmueller, J. (2011). Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies. *Political Analysis* 20, 25–46.
- Hastie, T., R. Tibshirani, and J. Friedman (2009). *Elements of Statistical Learning*. Springer.

- Hazlett, C. (2013). A balancing method to equalize multivariate densities and reduce bias without a specification search. *Working draft*.
- Heckman, J. J., H. Ichimura, and P. Todd (1998). Matching as an econometric evaluation estimator. *The Review of Economic Studies* 65(2), 261–294.
- Heckman, J. J., H. Ichimura, and P. E. Todd (1997). Matching as an econometric evaluation estimator: Evidence from evaluating a job training programme. *The Review of Economic Studies* 64(4), 605–654.
- Hirano, K. and G. Imbens (2001). Estimation of causal effects using propensity score weighting: An application to data on right heart catheterization. *Health Services and Outcomes Research Methodology* 2, 259–278.
- Hirano, K., G. W. Imbens, and G. Ridder (2003). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica* 71(4), 1161–1189.
- Hofmann, T., B. Schölkopf, and A. J. Smola (2008). Kernel methods in machine learning. *The Annals of Statistics*, 1171–1220.
- Imai, K. and M. Ratkovic (2014). Covariate balancing propensity score. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 76(1), 243–263.
- Imbens, G. W. (2004). Nonparametric estimation of average treatment effects under exogeneity: A review. *Review of Economics and Statistics* 86(1), 4–29.
- Imbens, G. W. and D. B. Rubin (2015). *Causal Inference for Statistics, Social, and Biomedical Sciences*. Cambridge University Press.
- Kang, J. D. and J. L. Schafer (2007). Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical Science* 22(4), 523–539.
- LaLonde, R. J. (1986). Evaluating the econometric evaluations of training programs with experimental data. *The American Economic Review*, 604–620.
- Lee, B. K., J. Lessler, and E. A. Stuart (2010). Improving propensity score weighting using machine learning. *Statistics in medicine* 29(3), 337–346.
- Lee, B. K., J. Lessler, and E. A. Stuart (2011). Weight trimming and propensity score weighting. *PloS one* 6(3), e18174.
- Lunceford, J. K. and M. Davidian (2004). Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. *Statistics in Medicine* 23(19), 2937–2960.
- McCaffrey, D. F., G. Ridgeway, and A. R. Morral (2004). Propensity score estimation with boosted regression for evaluating causal effects in observational studies. *Psychological Methods* 9(4), 403.
- McCullagh, P. and J. A. Nelder (1989). *Generalized linear models*. CRC press.
- Normand, S.-L. T., M. B. Landrum, E. Guadagnoli, J. Z. Ayanian, T. J. Ryan, P. D. Cleary, and B. J. McNeil (2001). Validating recommendations for coronary angiography following acute myocardial infarction in the elderly: a matched analysis using propensity scores. *Journal of clinical epidemiology* 54(4), 387–398.
- Robins, J., M. Sued, Q. Lei-Gomez, and A. Rotnitzky (2007). Comment: Performance of double-robust estimators when inverse probability weights are highly variable. *Statistical Science* 22(4), 544–559.
- Robins, J. M., A. Rotnitzky, and L. Zhao (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association* 89, 846–866.
- Robins, J. M. and N. Wang (2000). Inference for imputation estimators. *Biometrika* 87(1), 113–124.

- Rosenbaum, P. and D. Rubin (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika* 70(1), 41–55.
- Rosenbaum, P. and D. Rubin (1984). Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association* 79, 516–524.
- Rosenbaum, P. R. and D. B. Rubin (1985). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician* 39(1), 33–38.
- Rubin, D. B. (2009). Should observational studies be designed to allow lack of balance in covariate distributions across treatment groups? *Statistics in Medicine* 28(9), 1420–1423.
- Savage, L. J. (1971). Elicitation of personal probabilities and expectations. *Journal of the American Statistical Association* 66(336), 783–801.
- Selten, R. (1998). Axiomatic characterization of the quadratic scoring rule. *Experimental Economics* 1(1), 43–62.
- Smith, J. A. and P. E. Todd (2005). Does matching overcome lalonde’s critique of nonexperimental estimators? *Journal of econometrics* 125(1), 305–353.
- Stuart, E. A. (2010). Matching methods for causal inference: A review and a look forward. *Statistical Science* 25(1), 1–21.
- Wager, S. and S. Athey (2015). Estimation and inference of heterogeneous treatment effects using random forests. *arXiv preprint arXiv:1510.04342*.
- Wahba, G. (1990). *Spline models for observational data*, Volume 59. SIAM.
- Zhao, Q. and D. Percival (2015). Primal-dual Covariate Balance and Minimal Double Robustness via Entropy Balancing. *ArXiv e-prints* 1501.03571.